MODELO PARA LA PREDICCIÓN DE LA DESERCIÓN DE ESTUDIANTES DE PREGRADO, BASADO EN TÉCNICAS DE MINERÍA DE DATOS

ANIBAL JOSE CAMARGO GARCIA

UNIVERSIDAD DE LA COSTA - CUC

MAESTRIA EN INGENIERIA

BARRANQUILLA

2020

MODELO PARA LA PREDICCIÓN DE LA DESERCIÓN DE ESTUDIANTES DE PREGRADO, BASADO EN TÉCNICAS DE MINERÍA DE DATOS

ANIBAL JOSE CAMARGO GARCIA

Trabajo de grado para optar el Titulo de Magister en Ingeniería

Tutores

Ing. Emiro De la hoz Franco, PhD

Ing. Fabio Mendoza Palechor, PhD

UNIVERSIDAD DE LA COSTA - CUC

MAESTRIA EN INGENIERIA

BARRANQUILLA

2020

NOTA DE ACEPTACIÓN

_____

_____

_____

_____

Firma del Presidente del Jurado

_____

Jurado

_____

Jurado

Barranquilla, julio de 2020

Agradecimientos

Quiero agradecer a la Universidad de la Costa por brindarme la oportunidad de hacer parte de esta maravillosa institución y de apoyarme en la ejecución de este proyecto.

A mis tutores Emiro de la Hoz y Fabio Mendoza, los cuales me orientaron a aterrizar las ideas y aclarar todas las dudas.

A mis compañeros de la Maestría, por su apoyo, confianza, por compartir sus conocimientos y fortalecer los lazos de amistad.

A mi familia, por ser un pilar fundamental en todo este proceso y apoyarme en cada instante.

Resumen

El objetivo principal de este proyecto de investigación es crear un modelo para la predicción de la deserción de estudiantes de pregrado en la Universidad de la Costa - CUC, a partir del análisis de diferentes factores socioeconómicos y académicos. El estudio requirió de la ejecución de una serie de fases: caracterización, experimentación, desarrollo y evaluación. Durante la fase de caracterización se construyó un conjunto de datos (dataset), a partir de la compilación de los datos demográficos, culturales, sociales, familiares, educativos, estatus socioeconómico y perfil psicológico de cada estudiante, de los periodos comprendidos entre 2013-1 y 2018-2. Tal información fue recopilada a partir de los formatos de inscripción que diligencian los estudiantes cuando ingresan a la institución, un total de 1.606 registros únicos de estudiantes fueron recopilados. Durante la fase de experimentación se evaluaron distintas técnicas de aprendizaje automático (Machine Learning) de las categorías: redes bayesianas, máquinas de soporte vectorial y árboles de decisiones. El algoritmo con el cual se obtuvo la mejor tasa de aciertos fue Random forest (de la categoría árboles de decisión), con una exactitud del 84.8%. En la fase de desarrollo se integró el modelo a una aplicación que permite predecir si un estudiante o un grupo de ellos desertará o no. Por último, en la fase de evaluación se sometió la aplicación a diferentes tipos de pruebas para valorar tanto la funcionalidad de la interface gráfica con el usuario final como la tasa de aciertos en cuanto a la predicción de la deserción, los resultados han coincidido con la precisión obtenida en la fase experimental.

Palabras clave**:** educación superior**,** deserción, minería de datos, árboles de decisión, clasificación, predicción

Abstract

The main objective of this research project is to create a model for the prediction of undergraduate student desertion at the Universidad de la Costa - CUC, based on the analysis of different socioeconomic and academic factors. The study required the execution of a series of phases: characterization, experimentation, development and evaluation. During the characterization phase, a dataset was constructed, based on the compilation of demographic, cultural, social, family, educational, socioeconomic status and psychological profile data of each student, for the periods between 2013-1 and 2018-2. Such information was collected from the registration forms that students fill out when they enter the institution, a total of 1,606 unique student records were collected. During the experimental phase, different machine learning techniques were evaluated for the categories: Bayesian networks, support vector machines, and decision trees. The algorithm with which the best hit rate was obtained was Random forest (from the decision tree category), with an accuracy of 84.8%. In the development phase, the model was integrated into an application that allows us to predict whether a student or a group of students will drop out or not. Finally, in the evaluation phase, the application was subjected to different types of tests to evaluate both the functionality of the graphic interface with the final user and the success rate in terms of desertion prediction, the results have coincided with the precision obtained in the experimental phase.

***Keywords*:** higher education, dropout, data mining, decision tree, classification, prediction

Contenido

Lista de tablas y figuras

Tablas

Figuras

Introducción

Contexto

La deserción estudiantil en gran medida es la responsable de grandes desilusiones en los jóvenes que ingresan al sistema de educación superior dado que, por diversas razones no logran graduarse. Lo cual evidencia la ineficiencia del sistema de educación superior, al no proveer las herramientas que mitiguen las altas tasas de deserción, en procura de mantener la mayor cantidad posible de estudiantes que ingresan al sistema educativo, sin detrimento del proceso formativo. El flajelo de la deserción dificulta los procesos tendientes a ampliar la cobertura de la educación superior, lo cual a su vez incide en la postergación de la formación de capital humano de calidad en el país (Universidad de los Andes, 2014). Según (Martelo, Herrera, & Villabona, 2017), los factores que se han mayormente identificado, como de alta incidencia en la deserción estudiantil, se clasifican en: psicoeducativos, evolutivos, familiares, económicos, institucionales y sociales.

En Colombia según el Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior - SPADIES, con corte de marzo del 2017, el porcentaje de deserción por cohorte fue de 45,09%. Este porcentaje es muy alto comparado con el de España, donde la deserción es del 24,9%, la cifra es preocupante cuando se le compara con la tasa de deserción promedio de la Unión Europea que sólo alcanzó en dicho año el 12,8% (Universidad del Rosario, 2015). Sin embargo, en el contexto latinoamericano, la brecha no es tan grande si se compara con países como Argentina o Venezuela, 43% y 52% respectivamente (Universidad del Rosario, 2015). Las tasas de deserción de los países latinoamericanos son muy altas, pese a que en ellos se han planteado políticas tendientes a intervenir sobre determinados frentes de acción ante el problema, específicamente sobre factores de tipo económico. Sin embargo, en esta materia la oferta de créditos educativos no siempre contempla el desarrollo de estrategias de autogestión

con miras a cubrir los gastos correspondientes a los costos crediticios. Lo anterior conlleva a generar una baja capacidad de retención, por parte de algunas instituciones educativas, debido a sus particularidades directamente asociadas a su normatividad académica y a los procesos organizacionales dentro de tales instituciones (García Mendoza & Soto Cantero, 2014).

Para tratar de contrarrestar y prevenir los factores que influyen en la deserción estudiantil Universitaria, el Ministerio de Educación Nacional - MEN, creó el SPADIES. Sistema que permite medir y monitorear los factores determinantes de la deserción, conocer su evolución en el tiempo y ver cómo se comportan diferentes instituciones y regiones con relación a sus tasas de deserción (MEN & EAFIT, 2011). SPADIES facilita el seguimiento estadístico a los niveles de deserción (por programa académico, institución, sector, tipo de institución, región y áreas del conocimiento), con el propósito de poder analizar el comportamiento de factores determinantes del fenómeno, y en consecuencia facilitar la elección y evaluación de estrategias institucionales de apoyo a los estudiantes (MEN & EAFIT, 2011). El SPADIES "hace posible que cada institución cuente hoy con un perfil de sus estudiantes y con sistemas de alertas tempranas sobre los factores que los hacen vulnerables, lo que sirve para orientar de forma más eficiente apoyos y políticas" (MEN & EAFIT, 2011).

Problemática abordada y motivación

Basados en los lineamientos definidos por el MEN y a partir del análisis de la información extraída del SPADIES, la Corporación Universidad de la Costa - CUC, ha diseñado y operacionalizado diferentes estrategias y ha creado varios programas, todos ellos tendientes a promover la permanencia estudiantil, tales como: el Programa para el Acompañamiento y Seguimiento para la Permanencia Estudiantil - PASPE, el Programa de Consejería y Call Center, el curso de refuerzo académico y nivelación, el uso de Objetos Virtuales de Aprendizaje - OVA

en las asignaturas de mayores niveles de deserción, el fortalecimiento de los sistemas de información y la implementación del programa CUCJobs. Todos estos programas y estrategias han incidido positivamente en la disminución de los niveles de deserción estudiantil universitaria, gracias a la detección temprana de posibles riesgos.

Según entrevista realizada a Fragozo, Directora de Bienestar Estudiantil de la CUC, las anteriormente mencionadas acciones no son suficientes para disminuir significativamente las tasas de deserción estudiantil acorde a los requerimientos del MEN (Z. Fragozo, comunicación personal, 12 de diciembre de 2017). De allí nació la necesidad de utilizar un software basado en modelos estadísticos, denominado EDUSTAY, empleado por los consejeros estudiantiles, Bienestar Universitario y el Departamento de Estadística de la CUC, para realizar el registro de diferentes servicios prestados a estudiantes de la institución, tales servicios son: asesorías psicológicas, orientación vocacional, monitorias académicas, intermediación laboral, talleres de crecimiento personal y asesoría académica. Lo anterior, permite centralizar las variables y los factores objeto de análisis que favorecen la toma de decisiones en pro de mitigar el fenómeno de la deserción estudiantil (Combita & Morales, 2015).

EDUSTAY realiza una evaluación de conocimiento a los estudiantes, identificando las debilidades académicas al momento de integrar sus estudios a la educación superior, facilitando el proceso de seguimiento de las debilidades detectadas. Además, gestiona las entrevistas realizadas a los estudiantes en la etapa de admisión, con el fin de poder robustecer la información registrada, en la hoja de vida del estudiante. Sin embargo, si bien EDUSTAY facilita a los consejeros el registro de las variables que el sistema académico de la institución no contempla y pese a que cuenta con un módulo de reportes, la Directora de Bienestar Estudiantil (Z. Fragozo, comunicación personal, 12 de diciembre de 2017) y el Director del departamento de Estadística

(A. Castro, comunicación personal, 04 de enero de 2018) de la CUC, en respectivas entrevistas realizadas, manifiestan que la herramienta permite un análisis predictivo básico, evidenciando debilidades en cuanto a la identificación de la deserción con miras a mejorar los niveles de retención estudiantil, a partir del análisis de la información socioeconómica y académica de los estudiantes. Los directores de Bienestar Estudiantil y de Estadística, de la CUC, son conscientes de que el siguiente paso es la utilización de estrategias relacionadas con la minería de datos, para abordar el problema desde un enfoque predictivo, con el propósito de disminuir la deserción estudiantil universitaria, en coherencia con lo planteado por (Combita & Morales, 2015).

Las técnicas de minería de datos se utilizan para transformar datos brutos en información y posteriormente en conocimiento útil (Asif, Merceron, Ali, & Haider, 2017). Tales técnicas son aplicadas a diferentes contextos o sectores (salud, gobierno y educación, entre muchos otros). Los autores (Asif et al., 2017) analizan varios casos donde fueron aplicadas técnicas de minería de datos, a continuación, se sintetizan:

- Desarrollo de modelos para predecir el rendimiento de los estudiantes en la universidad, basados en la información personal de los estudiantes. Los autores analizaron datos de 10.330 estudiantes del sector educativo búlgaro. Se emplearon 20 atributos que incluían género, año y lugar de nacimiento, lugar de residencia, y país, lugar y puntuación total de la educación anterior, el semestre actual y la puntuación total de la universidad. Se aplicaron algoritmos de minería de datos, como árbol de decisión C4.5, redes bayesianas, vecinos más cercanos – (KNN) y algoritmos de aprendizaje de reglas. El clasificador del árbol de decisión tuvo el mejor desempeño con la mayor precisión general, seguido del clasificador del aprendiz de reglas (JRip) y el clasificador k-NN. Sin embargo, todos los

clasificadores se desempeñaron con una precisión global inferior al 70% (Kabakchieva et al., 2011).

- Predicción de la probabilidad de éxito / fracaso en la universidad. Indicó que en el Examen de Certificado de Ingreso de Educación Superior Etíope – EHEECE (por su sigla en inglés) los principales factores que afectaron el rendimiento del estudiante fueron: el género, número de estudiantes en una clase, número de cursos dados en un semestre y el campo de estudio. En esta investigación se obtuvo una tasa de precisión de la predicción de un 92.34% (Yehuala, 2015).
- En (Oskouei & Askari, 2014) analizaron el rendimiento académico de los estudiantes de la escuela secundaria y de licenciatura en el Irán. Consideraron los datos de 500 estudiantes con nivel de secundaria y 600 estudiantes con nivel de licenciatura. Aplicaron varios clasificadores como Naive Bayes, árbol de decisión C4.5, Bosque aleatorio y redes neuronales, y metaclasificadores como Bagging, Boosting o Adaboost, para clasificar a los estudiantes en 2 clases: Aprobar, reprobar. Los resultados revelaron que características como el nivel educativo de los padres, los resultados de exámenes anteriores y el género influyen en la predicción. La mejor precisión del 96% se obtuvo con el árbol de decisión C4.5.

Los anteriores casos fueron plasmados en el estudio realizado por (Asif et al., 2017), analizando situaciones reales en distintas Universidades y los resultados obtenidos al aplicar algoritmos de minería de datos en el ámbito educativo.

El Director (A. Castro, comunicación personal, 04 de enero de 2018) y los Auxiliares de dicho departamento, han manifestado que el análisis estadístico (predictivo) generado por EDUSTAY no es eficiente, debido a que es difícil de actualizar, por la dinámica y contexto

cambiante de los estudiantes año tras año. Por ello, se hace necesario el planteamiento de una solución que permita predecir de manera eficiente la deserción de estudiantes de pregrado, basada en técnicas de Minería de Datos y mejorando, por tanto, los niveles de retención estudiantil en la CUC. A partir de lo anteriormente mencionado, nace el siguiente interrogante: ¿Qué herramienta tecnológica, basada en técnicas de Minería de Datos, permitiría mejorar los niveles de deserción estudiantil a nivel de pregrado en la Universidad de la Costa - CUC?

Justificación

La Corporación Universidad de la Costa – CUC, como Institución de Educación Superior – IES, domiciliada en la ciudad de Barranquilla - Colombia, ofrece diversos programas académicos en procura de formar profesionales y con el fin de promover el desarrollo socioeconómico a nivel local y regional, lo que exige una constante renovación y revisión de todos los conceptos inherentes al ámbito competitivo y de calidad, en coherencia con la globalización.

Desde el punto de vista de la relevancia, se determina importante porque al analizar las tasas de deserción de la CUC, se observa una reducción. En el periodo del 2006-1, la tasa de deserción fue del 28.91%. En el 2007-1, se notó una disminución a 22.21%, esta constante se mantuvo en los periodos siguientes, reduciendo significativamente, hasta llegar al 10.35% en el periodo 2018-2. Para una población estudiantil de 12 mil estudiantes que maneja de Universidad, esta cifra de deserción representa un alto costo, y detectar a tiempo los estudiantes en riesgo, ayudaría a tomar acciones oportunas y disminuir la tasa.

En cuanto a los referentes legales y de alcance, se afirma aplicable ya que las disposiciones de ley emanadas por el MEN, se enfocan en la reestructuración y modernización de todos los programas académicos en el país, entre estos la piscología, a fin

de propender por un escenario futuro donde los profesionales tengan características y niveles de desempeño muy superiores danto apertura a un panorama más universal y productivo (Guzmán Ruiz et al., 2009; Ministerio de Educación de Colombia, 2006).

Objetivos

El presente estudio plantea como objetivo general: Desarrollar una aplicación prototipo que permita predecir la deserción estudiantil a nivel de pregrado en la Universidad de la Costa - CUC, a partir del análisis de información socioeconómica y académica de los estudiantes de pregrado, mediante la implementación de un modelo funcional basado en técnicas de Aprendizaje Automático.

Para la consecución del objetivo general anteriormente mencionado, se han planteado los siguientes tres objetivos específicos en coherencia con la ejecución de las fases de caracterización (primer objetivo), experimentación (segundo objetivo), desarrollo y evaluación (tercer objetivo).

- Identificar y documentar las variables socioeconómicas y académicas de los estudiantes, y las técnicas de Minería de Datos, como apoyo a la construcción del conjunto de datos.
- Simular escenarios de experimentación para identificar los componentes del modelo predictivo funcional.
- Desarrollar el aplicativo prototipo, que soporte el modelo predictivo funcional, evaluando su nivel de la calidad en cuanto a la tasa de aciertos.

Mapa del documento

En el capítulo primero se introduce al lector sobre el contexto objeto de estudio, la problemática o motivación, las razones que justifican el abordaje de esta, los objetivos trazados y una descripción sintética los capítulos que la constituyen.

En el capítulo segundo se abordan los ejes temáticos que fundamentan la investigación, en relación a: deserción en instituciones de educación superior, deserción como comportamiento individual, deserción desde el punto de vista institucional, deserción según la perspectiva estatal o nacional, el Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior - SPADIES, factores determinantes de la deserción, técnicas para la predicción utilizadas en procesos de minería de datos, metodologías del proceso de minería de datos, técnicas basadas en aprendizaje supervisado.

En el tercer capítulo se abordan las investigaciones relacionadas con la investigación, agrupando en cuatro tipos: Análisis de satisfacción, predicción del rendimiento estudiantil, riesgo de reprobación y análisis de retención. Así como los clasificadores y métricas utilizadas en dichas investigaciones y los problemas comunes.

En el cuarto capítulo se detalla el proceso para la creación del modelo para predecir la deserción de los estudiantes. Seleccionando la metodología aplicada en el proyecto de investigación, el proceso de recopilación de los datos, la etapa de preprocesamiento y la simulación de los escenarios aplicando técnicas de aprendizaje automático.

En el quinto capítulo se detalla el desarrollo de la aplicación prototipo, la arquitectura, la descripción funcional y el paso a paso, uso de la predicción masiva y las mejoras futuras.

Fundamentación conceptual

El presente estudio se fundamenta a partir de los ejes conceptuales deserción en instituciones de educación superior y técnicas utilizadas para la predicción, en procesos de minería de datos. En este capítulo se hace un análisis detallado de cada uno de estos ejes. En cuanto al primero, inicialmente se conceptualiza la deserción estudiantil tanto desde sus causas como desde el enfoque individual, institucional y estatal. Adicionalmente, se analizan los componentes funcionales del SPADIES y su contribución como apuesta de país, a la disminución de los niveles de deserción, la descripción de este eje culmina con el análisis de los factores determinantes que inciden en la deserción estudiantil. En cuanto al segundo eje, en primera instancia es necesario analizar las metodologías más ampliamente utilizadas en los procesos inherentes a la minería de datos, luego se revisan varias categorías de algoritmos basados en técnicas de aprendizaje supervisado, que arrojan mejores resultados al ser utilizados en procesos de minería de datos, con miras a la predicción. El capítulo culmina con una revisión conceptual de las métricas de calidad más preponderantes, al momento de evaluar la capacidad de acierto de los modelos implementados, a partir de las técnicas previamente descritas para procesos de predicción.

Deserción en instituciones de educación superior

La deserción estudiantil en instituciones de educación superior inicialmente requiere de una definición formal, resaltando que, si bien no existe consenso, se evidencian unos componentes comunes que demarcan un eje central definitorio. La deserción puede ser abordada desde los comportamientos particulares de los individuos y cómo éstos inciden en ella, hasta una visión más macro, basada en las generalidades propias de la visión colectiva institucional y

nacional. Para ampliar la perspectiva, respecto a este tema, es necesario ver las acciones a nivel gubernamental, que se han tomado para mitigar el flagelo de la deserción y los factores que históricamente la han determinado.

Deserción Estudiantil

En (Vicent Tinto, 1989) se define la deserción estudiantil como la frustración para finalizar un curso especifico de acción o lograr un fin determinado, en busca de la cual el sujeto ingresó a una determinada institución de educación superior. Adicionalmente, el autor plantea que la deserción depende de las intenciones individuales, sociales e intelectuales, por las cuales las personas elaboran metas deseadas en una cierta universidad. En otro estudio (Vincent Tinto, 1975) analiza el proceso de deserción escolar, distinguiendo los tipos de conducta propios del abandono escolar, enfocándose en el retiro académico y el retiro voluntario. El autor plantea que se debe no sólo a que estos comportamientos involucran a diferentes personas, además indica que son el resultado de diferentes patrones de interacción dentro del entorno universitario. Por lo tanto, el autor plantea, que aunque el retiro académico está más estrechamente asociado con el rendimiento académico, la deserción en la forma de retiro voluntario no lo está. Esta última forma de retiro parece estar más relacionada con la falta de congruencia entre el individuo y el clima intelectual de la institución o el sistema social, que integra a sus pares (Vincent Tinto, 1975). El autor plantea que la deserción estudiantil puede analizarse desde varias aristas y que se basa en los diferentes tipos de abandono. Según él, existen tres puntos de vista, dependiendo de las partes involucradas e interesadas en el proceso: la deserción como comportamiento individual, la deserción desde el punto de vista institucional y la deserción según la perspectiva estatal o nacional.

Por otra parte, en (Castaño et al., 2004) definen la deserción como un momento al que un estudiante se enfrenta, cuando anhela y no logra culminar su plan de estudio, considerando como desertor a aquella persona que no muestra actividad académica a lo largo de tres semestres (o períodos académicos) seguidos.

Deserción como Comportamiento Individual

Según (Vicent Tinto, 1989), la idea fundamental para dar una enunciación de deserción, apropiada a la representación de la persona, es el entendimiento de que las connotaciones que un estudiante confiere a su conducta pueden diferenciarse de las que un espectador asigna a la misma actuación. El autor argumenta que el hecho de dejar una universidad puede poseer conceptos variados y totalmente diversos para aquel que es partícipe o son perjudicados por esa conducta. No obstante un espectador, logra determinar el abandono como un revés en concluir un plan de estudios, los estudiantes pueden deducir su renuncia como un camino provechoso en dirección a la obtención de un objetivo (Vicent Tinto, 1989).

Desde la perspectiva individual, la deserción significa el fracaso para finalizar un curso especifico u obtener un logro deseado, en busca de la cual la persona se incorporó a una determinada institución de educación superior (Vicent Tinto, 1989).

Deserción desde el punto de vista Institucional

Según (Vicent Tinto, 1989), definir la deserción mediante un enfoque institucional es, en ciertos aspectos, un trabajo más sencillo que hacerlo conforme con el punto de vista individual. En otros, a pesar de lo cual, es ampliamente más difícil. El autor plantea que es más simple en la manera de que todas las personas que desertan una institución de educación superior logran, valorar las razones expuestas para lograr, ser catalogados como desertores. Los estudiantes que desertan crean un puesto en los estudiantes que pudieron ocupar dicho espacio. En consecuencia,

el autor argumenta que esto genera problemas financieros al generar un desequilibrio en los ingresos de las instituciones. Se evidencia más en el sector privado, en donde las matrículas forman parte fundamental de los ingresos, sin embargo, afecta al sector público, debido a la falta escasa de presupuesto (Vicent Tinto, 1989).

Deserción según la perspectiva estatal o nacional

En (Vicent Tinto, 1989) indican que es diferente cuando la perspectiva de la deserción es estatal. Por ejemplo, la dimisión que se realiza entre instituciones de educación públicas no puede representar deserciones en el sentido estricto de la palabra, debido a que son cambios interiores realizados en el sector estatal. No obstante, si se genera una salida de estudiantes con destino a las instituciones privadas, existe la probabilidad que las dimisiones o abandonos sean tenidos en cuenta como deserciones. En este sentido, solo las personas que abandonan de todo el sistema de educación superior serán posiblemente consideradas como deserciones.

El gobierno nacional, por medio del ministerio de educación, abordó la deserción estudiantil en las instituciones de educación superior, en el Plan de Desarrollo denominado “La Revolución Educativa 2002-2006” (Ministerio Nacional de Educación, 2008). El plan de desarrollo consistió en establecer una estrategia del sector educativo para mejorar su calidad, cobertura y eficiencia. La deserción estudiantil involucra los objetivos institucionales y sectoriales de dichas políticas, igualmente de comprometer desperdicio de capital privado por parte de las familias y de distintos elementos educativos (Ministerio Nacional de Educación, 2008). Es importante agregar que el estado colombiano ha generado una herramienta para el estudio y tratamiento de la deserción estudiantil denominada Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior –SPADIES– (Ministerio Nacional de Educación, 2008).

Spadies

El Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior - SPADIES, reúne y clasifica información que faculta realizar una búsqueda a las circunstancias socioeconómicas y académicas de los estudiantes que se han incorporado a la educación superior en el país. De este modo, se puede permite saber la evolución y el estado del provecho y la caracterización académica de los estudiantes, esto es beneficioso para constituir los factores definitivos de la deserción, para evaluar el riesgo de deserción de los estudiantes y plantear y desarrollar las actividades de refuerzo, dirigidas a promover su continuidad y culminación. El SPADIES pertenece al Sistema Nacional de Información de la Educación Superior —SNIES y logra inferir como una parte propia de este, concentrado a la búsqueda especifica de un fenómeno de singular atractivo en el sector educativo, como es la deserción estudiantil (Ministerio de Educación de Colombia, 2019). El seguimiento que se realiza en el SPADIES permite conocer los siguientes aspectos (Ministerio Nacional de Educación, 2008):

- Cantidad de desertores y estimaciones de deserción
- El rasgo de cada estudiante: cualidades socioeconómicas, académicas individuales
- Las razones o componentes claves de la deserción
- Los datos para la valoración de rendimiento y retroalimentación de actividades llevadas a cabo para reducir la deserción estudiantil
- Valoración del riesgo de deserción de los estudiantes

Factores determinantes de la deserción

En (Castaño et al., 2004), realizan un análisis de varios autores y las distintas perspectivas de la deserción, agrupándolas en cuatro factores determinantes:

- Factores individuales: edad, género, estado civil, entorno familiar, calamidad y problemas de salud, integración social, incompatibilidad horaria con actividades extraacadémicas y expectativas no satisfechas.
- Factores académicos: orientación profesional, tipo de colegio, rendimiento académico, calidad del programa, método de estudio, resultado en el examen de ingreso, insatisfacción con el programa u otros factores académicos y números de materias.
- Factores institucionales: normalidad académica, becas y formas de financiamiento, recursos universitarios, orden público, entorno político, nivel de interacción personal con los profesores y estudiantes.
- Factores socioeconómicos: estrato, situación laboral, situación laboral de los padres e ingresos, dependencia económica, personas a cargo, nivel educativo de los padres y entorno macroeconómico del país.

Técnicas para la predicción utilizadas en procesos de minería de datos

Antes de analizar las técnicas utilizadas para la construcción de modelos predictivos, en procesos de minería de datos, es necesario comprender tanto la definición del término como las metodologías empleadas para los procesos de tratamiento de datos. Si bien existen dos grandes enfoques para abordar este tipo de estudios, como son: el enfoque orientado a los datos (*Data-Driven Approaches* - DDA) y el enfoque orientado al conocimiento (*Knowledge-Driven Approaches* – KDA). Por la tiplogía de estudio, este proyecto se ha decantado por el uso del primer enfoque. El DDA contempla diferentes categorias, como son: análisis de umbral (*Threshold Analysis*), métodos de aprendizaje automático (*Machine Learning Methods* – MLM), métodos de regresión (*Regression Methods*) y técnicas de correspondencia de curvas (*Curve Matching Techniques*). Luego de un exhaustivo análisis de la literatura científica, se ha

evidenciado que la categoría MLM es de las más utilizadas para procesos de clasificación con miras a la predicción. A su vez, esta categoría contempla una serie de subcategoría de métodos, tales como: clasificadores bayesianos (*Bayesian Classifiers* - BC), clasificadores basados en instancias (*Instance Based Classifiers*), máquinas de soporte vectorial (*Support-Vector Machine* – SVM), árboles de decisión (*Decision Trees* – DT), redes neuronales artificiales (*Artificial Neural Networks*), lógica difusa (*Fuzzy Logic*) y modelos de markov (*Markov Models*), entre otras. Daba la amplia diversidad de subcategorías y de técnicas propias de cada una de éstas, el estudio se basó concretamente en las subcategorías: DT, SVM y BC, cada una de ellas son descritas en el cuerpo de esta sección. Posteriormente se conceptualizan las métricas que permiten evaluar la calidad de los modelos generados a partir de la implementación de las técnicas.

Minería de datos

Si bien no existe una única definición del término minería de datos, las conceptualizaciones propuestas por (Aggarwal, 2015), (Brown, 2014), (Han et al., 2012), (Zaki & Meira, 2013) y (M. Perez, 2014) dan una clara evidencia de su importancia en la actualidad, como herramienta para resolver gran variedad de problemáticas en las cuales se requiera el tratamiento de considerables y diversos capacidad de datos, para la identificación de relaciones no triviales en los mismos. A continuación, un análisis detallado de cada una de estas definiciones.

En (Aggarwal, 2015) definen la minería de datos como el estudio de la recolección, limpieza, procesamiento, análisis y adquisición de información útil desde los datos. Existe una amplia variación en cuanto a los dominios de problemas, aplicaciones, formulaciones y

representaciones de datos que se encuentran en las aplicaciones reales. Por lo tanto, "minería de datos" se emplea para explicar estos distintos aspectos del procesamiento de datos.

En (Brown, 2014) la definen como la forma en que los empresarios pueden explorar los datos de forma independiente, hacer descubrimientos informativos y poner esa información a trabajar en el día a día de la empresa.

Según (Han et al., 2012), “la minería de datos es el proceso de descubrir patrones y conocimientos interesantes a partir de grandes cantidades de datos. Las fuentes de datos pueden incluir bases de datos, almacenes de datos, la Web, otros repositorios de información o datos que se transmiten dinámicamente al sistema”.

En (Zaki & Meira, 2013) la definen como el proceso de descubrir patrones perspicaces, interesantes y novedosos, así como modelos descriptivos, comprensibles y predictivos a partir de datos a gran escala.

En (M. Perez, 2014) se define minería de datos como como el conjunto de técnicas orientadas a descubrir información almacenada en grandes conjuntos de datos. Analizando patrones, conductas, preferencias, agrupaciones y propiedad del entendimiento que ofrecen los datos. En la actualidad se cuenta con gran cantidad de datos, haciéndose necesario, poder estudiarlos minuciosamente para sacar de una forma automatizada el conocimiento incluido en ellos, empleando metodologías especializadas respaldadas en herramientas tecnológicas.

La avalancha de datos es el resultado directo de los avances tecnológicos y de la informatización que se recaba a partir de todos los aspectos de la vida moderna. Por lo tanto, es natural examinar si se puede extraer información concisa y posiblemente procesable de los datos disponibles para objetivos específicos de aplicación. Aquí es donde entra en juego la tarea de la minería de datos. Los datos en bruto pueden ser arbitrarios, no estructurados, o incluso en un

formato que no es inmediatamente adecuado para el procesamiento automatizado. Por ejemplo, los datos recogidos manualmente pueden extraerse de fuentes heterogéneas en diferentes formatos y posteriormente pueden ser procesados por un programa informático automatizado, para obtener información relevante (Aggarwal, 2015).

Metodologías propias del proceso de minería de datos

La comunidad científica ha definido varias metodologías para la implementación de procesos basados en minería de datos, las cuales han sido ampliamente validadas. A continuación se conceptualizan las más utilizadas: 1) descubrimiento de conocimiento en bases de datos (*Knowledge Discovery in Databases* – KDD), 2) muestreo, exploración, modificación, modelado y evaluación (*Sample, Explore, Modify, Model and Assess* – SEMMA) y 3) proceso estándar de la industria cruzada para la minería de datos (*Cross Industry Standard Process for Data Mining* - CRIPS-DM).

Descubrimiento de conocimiento en bases de datos – KDD.

En (Timarán Pereira et al., 2016) definen el KDD como un proceso automático en el que se componen dos etapas, descubrimiento y análisis. Radica en obtener patrones en modo de reglas o funciones, para que el usuario analice los datos. Dicha actividad acarrea regularmente preprocesar los datos, minería de datos y mostrar los resultados. El KDD debe verse como un paso esencial en el procedimiento de descubrimiento de conocimiento, tal como se muestra en la **Figura 1.** Dicha figura debe entenderse como una secuencia iterativa de los siguientes pasos, definidos en (Han et al., 2012), es decir: selección de datos, preprocesamiento / limpieza, transformación / reducción, minería de datos e interpretación / evaluación.

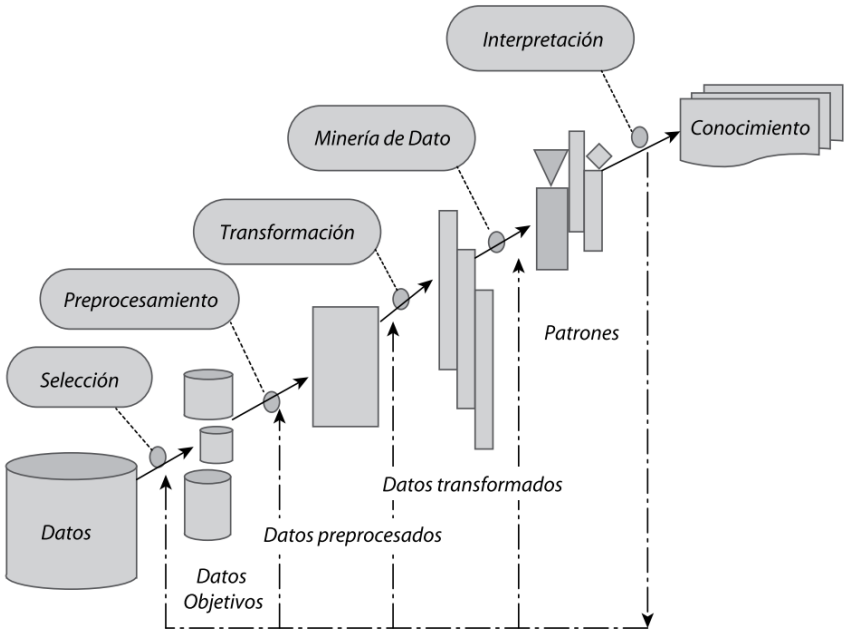


Figura 1. Proceso KDD.

Tomada de (Timarán Pereira et al., 2016)

En cuanto a la etapa de **selección de los datos**, el propósito principal es determinar el conocimiento importante y preexistente, definiendo los objetivos del proceso KDD. Seleccionar los datos cambia conforme con las metas del negocio (Timarán Pereira et al., 2016). En este paso se decidirá qué parte de los datos se va a utilizar para la minería de datos. El resultado de esta tarea es la justificación de la inclusión y la exclusión. Se darán las razones para incluir o excluir cada parte de los datos que se tienen, basándose en la relevancia para sus objetivos, la calidad de los datos y los problemas técnicos, como los límites del número de campos o filas que pueden manejar sus herramientas o la idoneidad de los formatos de datos para sus necesidades (Brown, 2014).

En cuanto al **preprocesamiento-limpieza**, los datos seleccionados en el paso anterior probable que no estén perfectamente limpios (sin errores). Por lo tanto, se deberán realizar cambios, probablemente rastreando fuentes para hacer correcciones específicas de datos, excluyendo algunos casos o celdas individuales (elementos de datos), o reemplazando algunos

elementos de datos con valores predeterminados o reemplazos seleccionados mediante una técnica de modelado más sofisticada. Se puede optar por utilizar sólo subconjuntos de los datos para todo o parte de su trabajo de minería de datos (Brown, 2014). El resultado de esta labor es la parte de la limpieza de datos, que documenta con detalle, todas las decisiones y acciones utilizadas para limpiar los datos. Este informe debe cubrir y referirse a cada problema de calidad de datos, que se identificó en la actividad de verificación de la calidad de datos, en la etapa de comprensión de datos del proceso. El informe también debe abordar el impacto potencial en los resultados de las elecciones que han realizado durante la limpieza de datos (Brown, 2014). Las acciones realizadas ante~~s~~ datos con errores o anómalos, pueden ser (Beltran, 2016): ignorarlos, filtrarlos (eliminando o reemplazando una columna), filtrado de filas, reemplazo de valores (por el valor "nulo", máximos, mínimos o medias) o discretización (transformar los valores continuos por discretos).

Mediante la **transformación-reducción** de datos, "se buscan características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos" (Timarán Pereira et al., 2016).

En cuanto al proceso de aplicación de técnicas de **minería de datos** para producir prototipos o modelos que pueden ser descriptivos o predictivos, tales prototipos tratan de tasar valores futuros o no usal de las variables de interés, llamadas variables objetivo, clases o dependientes, con la utilización de otras variables llamadas predictivas o independientes. Dentro de las labores predictivas se encuentran la clasificación y la regresión. Los prototipos o modelos descriptivos reconocen patrones que aclaran o demuestran los datos; su funcionalidad es para

analizar las propiedades de los datos observados, no para la predicción de datos nuevos, tales como la identificación de grupos de personas con estilos o gustos afín o detectar comportamientos de compra de clientes en una particular zona de algún lugar. Dentro de las actividades descriptivas se enumeran los patrones secuenciales, reglas de asociación, técnicas de agrupamiento o clustering y análisis de correlaciones (Timarán Pereira et al., 2016). La selección del algoritmo de minería de datos comprende la elección de los métodos por suministrar en la indagación de patrones en los datos, así como la determinación sobre los prototipos y variables más conveniente, dependiendo del tipo de datos (numéricos o categóricos) por emplear (Timarán Pereira et al., 2016).

En cuanto al proceso de **interpretación-evaluación**, como su nombre lo indica se evalúan las métricas de calidad luego de aplicar las técnicas seleccionadas en los pasos anteriores. Existe un variado repertorio de métricas y la interpretación de los resultados con miras a la identificación de cuáles son las técnicas más apropiadas, dependerá en gran medida de: la tipología de estudio (descriptivo, predictivo o prospectivo), la tipología de técnicas (basadas en aprendizaje supervisado o no supervisado) y de las particularidades propias de los criterios de clase (biclase o multiclase).

Muestreo, exploración, modificación, modelado y evaluación – SEMMA.

La metodología SEMMA ofrece un procedimiento sencillo de comprender, posibilitando un desarrollo y mantenimiento organizado y adecuado de los proyectos. Otorga una estructura lógica para su generación, formación y transformación, apoyando al presentar soluciones a variedad de problemas empresariales, enfocándose en los objetivos de negocio a partir de la aplicación de técnicas propias del proceso de minería de datos (Azevedo & Santos, 2008). SAS

Institute es un proyecto de la Universidad de Carolina del Norte (CITAR) que ha definido los cinco (5) pasos requeridos para la implementación de la metodología SEMMA, tales pasos son:

- **Muestreo** (*sample*): consiste en extraer una porción de un conjunto de datos, lo suficientemente grande como para contener información significativa que represente a la totalidad de los datos, pero lo suficientemente pequeña como para manipularlos rápidamente.
- **Exploración** (*explore*): consiste en la búsqueda de tendencias y anomalías imprevistas, con el fin de adquirir una mayor comprensión o entendimiento, en relación al significado de los datos.
- **Modificación** (*modify*): se basa en los cambios de los datos, a través la selección, creación, y transformación de las variables, que permitan el posterior proceso de selección del modelo.
- **Modelado (***model*): consiste en modelar los datos, posibilitando que el programa examine automáticamente una mezcla de datos, que pronostique de forma fiable el resultado deseado.
- **Evaluación (***asess*): consiste en evaluar los datos a partir del análisis de la utilidad y fiabilidad de los resultados.

Proceso estándar intersectorial para la minería de datos - CRIPS-DM.

Es un marco para interpretar los inconvenientes corporativo en tareas de minería de datos y efectuar proyectos de minería de datos, independientes como del área de aplicación y de la tecnología empleada (Huber et al., 2019). La metodología se fundamenta en prácticas reales sobre cómo las personas realizan proyectos (Timarán Pereira et al., 2016). CRIPS-DM consta de un grupo de tareas que se están estructuradas en cuatro (4) niveles: fases, tareas especializadas, tareas generales e instancias de proceso (ver Figura 2).

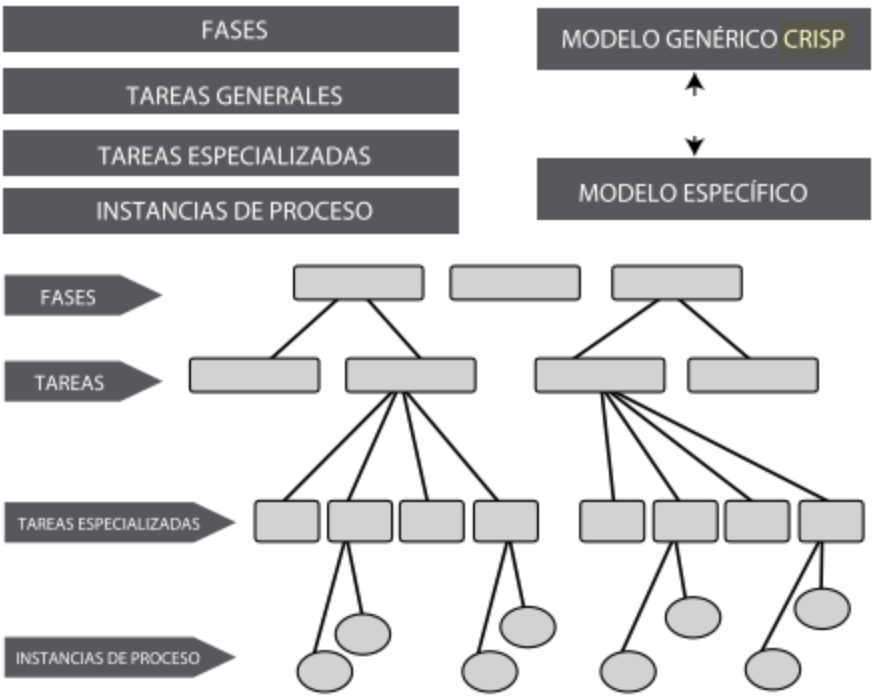


Figura 2. Fases CRIPS-DM.

Tomada de (Timarán Pereira et al., 2016)

En la parte superior, el proceso de minería de datos se estructura en un número reducido de fases. Cada etapa consiste en varias actividades comunes de segundo nivel. Este nivel se llama genérico, ya que pretende ser lo suficientemente frecuente para cobijar las posibles condiciones del proceso de minería de datos. Las tareas genéricas están diseñadas para ser lo más completas y constante posible. Estable denota que se desea que el modelo sea válido para desarrollos imprevistos basados en nuevas técnicas de modelado (Wirth, 2000).

En el nivel de tareas especializadas, se describe cómo deben llevarse a cabo los hechos en las tareas comunes y en situaciones particulares. Por ejemplo, en el nivel segundo hay una tarea llamada modelo de construcción. En el nivel tercero, se podría tener una tarea denominada construir un modelo de respuesta que contenga actividades específicas al problema y a la herramienta de minería de datos elegida. El detalle de fases y tareas, como pasos moderados

ejecutados en un orden particular, reproduce un orden idealizado de hechos. En la realidad, varias de las tareas se logran efectuar en una distribución distinta y frecuentemente será necesario regresar a ejercicios previos y duplicar ciertas actividades. El proceso CRISP-DM no trata de captar todas estas probables destinos por medio de la minería de datos, porque esto demandaría un modelo de proceso excesivamente complicado y las utilidades esperadas serían muy bajos (Wirth, 2000).

El nivel de instancia de proceso contempla la anotación de actividades, decisiones y frutos pertenecientes de la minería de datos. Una solicitud de proceso se dispone de acuerdo con las tareas determinadas en los niveles más altos, pero simboliza lo que efectivamente pasó en una obligación en particular, en vez de lo que sucede en general (Wirth, 2000).

Técnicas basadas en aprendizaje supervisado

Existen las técnicas basadas en aprendizaje supervisado y no supervisado. La primera trabaja con datos etiquetados de un conjunto de datos (Han et al., 2012), y en las no supervisadas no se tienen los datos etiquetados. Con esta última técnica, se utiliza para Agrupamiento (*Clustering*) que buscan datos basado en similitudes. Dabo que el conjunto de datos usado en el proyecto de investigación contenía datos etiquetados, se optó por la utilización de técnicas basadas en aprendizaje supervisado.

Las técnicas propias de los procesos de minería de datos admiten la creación de modelos predictivos y/o descriptivos. Según (Beltran, 2016), un modelo predictivo contesta interrogantes sobre datos futuros, todo lo contrario, un modelo descriptivo concede información sobre las correspondencia entre los datos y sus atributos.

Cuando se trata de hacer predicciones, lo que se busca es deducir algún tipo de resultado definido. Por ejemplo, si se está pensando en el clima de hoy, es posible que se desee saber si

lloverá (sí o no), un pronóstico del tiempo es un tipo de predicción. Los modelos utilizados para el pronóstico del tiempo se basan en la generación de proyecciones a partir de referentes históricos. Para este ejemplo, el resultado de cada día está bien definido (llovió o no llovió). Los modelos utilizados para predecir comparan los días en que llovió, para identificar las diferencias entre los dos y predecir lo que sucederá en el futuro. Situaciones como ésta, en las que las agrupaciones (valores de clases) están claramente definidas por algún resultado conocido, son aplicaciones para el aprendizaje supervisado. En los procesos de clasificación, que hacen uso de técnicas basadas en aprendizaje supervisado, se crean agrupaciones a partir de algún resultado conocido (o valor de una variable) (Brown, 2014).

En (Saravanan & Sujatha, 2018) plantean que los algoritmos que utilizan técnicas de aprendizaje supervisado necesitan que los humanos den la entrada y salida requerida, además de suministrar información sobre la precisión en el proceso de entrenamiento. Los autores indican que las técnicas de aprendizaje automático supervisado utilizan a partir de lo que se ha adquirido de conocimientos de datos anteriores y actuales con la ayuda de etiquetas para pronosticar los eventos.

Las técnicas basadas en aprendizaje supervisado se dividen en clasificadores probabilísticos y clasificadores lineales (Aboubakar et al., 2019). En los clasificadores probabilísticos encontramos clasificadores bayesianos (Naive Bayes, Redes Bayesianas) (Saravanan & Sujatha, 2018). En los lineales se encuentran las Maquinas de Soporte Vectorial (SMV), Regresión Logística (LR), Arboles de Decisiones (DT), Redes Neuronales (NN). Se profundizará sobre los clasificadores utilizados en el proyecto de investigación.

Arboles de Decisión

En (Quinlan, 1986) definen los Árboles de Decisión como métodos de aproximación a las funciones de valor discreto que es robusto a los datos ruidosos y capaz de aprender expresiones disyuntivas. Se construyen comenzando con la raíz del árbol y procediendo hasta sus hojas. Los métodos de árbol de decisión están diseñados para seleccionar un conjunto de variables de predicción y dividir sucesivamente un conjunto de datos en subgrupos a fin de mejorar la predicción (clasificación) de una variable objetivo (dependiente) (Quinlan, 1986; Veitch, 2004). A continuación, un ejemplo de un árbol de decisión:

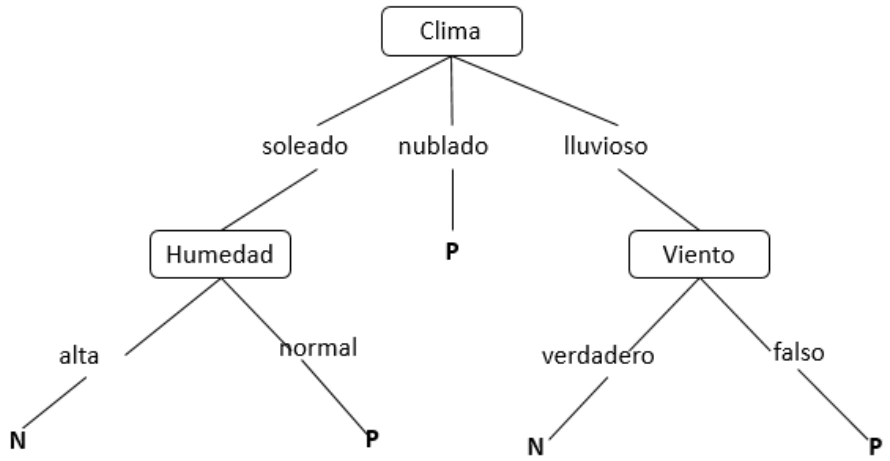


Figura 3. Árbol de decisión

Fuente: Elaboración propia

Los árboles de decisiones aproximan las funciones objetivo de valor discreto, en el que la función aprendida se representa mediante un árbol de decisión. Los árboles aprendidos también pueden representarse como conjuntos de reglas de "si" o "entonces" para mejorar la legibilidad humana (Mitchell, 1997).

Los árboles de decisión son a veces más interpretables que otros clasificadores como las redes neuronales y las máquinas de soporte vectorial, porque combinan de manera comprensible preguntas sencillas sobre los datos (Kingsford & Salzberg, 2008).

Los datos manejados en los árboles son representados con tres elementos diferentes: Atributos/Predictores, Valores y Clases. Existen dos tipos de árboles de decisión: árboles de clasificación y árboles de regresión. Los árboles de clasificación cuando las variables son discretas, predice el valor de una variable categórica mediante la creación de un modelo basado en uno o más atributos (Quinlan, 1986).

El uso de los árboles de decisiones se ha aplicado con éxito a una vasta serie de tareas que van desde el aprendizaje para diagnosticar casos médicos hasta el aprendizaje para evaluar el riesgo crediticio de los solicitantes de préstamos, deserción estudiantil (Kalles & Pierrakeas, 2006b; Mitchell, 1997; Pérez et al., 2019).

Los árboles de decisiones presentan varias ventajas para su uso, dentro de las cuales se encuentra: asisten a adquirir la elección más idónea desde una perspectiva probabilística en presencia de una serie de posibles decisiones, fácil combinación con herramientas para la toma de decisiones, útiles con o sin datos certeros y cualquier dato requiere una preparación mínima, tiende a ser preciso independientemente de si viola las suposiciones de los datos de origen (Beaulac & Rosenthal, 2019; Kingsford & Salzberg, 2008). Así mismo, presenta varias desventajes, dentro de las se encuentra: necesidad de uso de gran cantidad de datos, dificultad para obtención de la creación de un árbol decisiones óptimo, los cálculos pueden volverse complejos al enfrentarse con la carencia de certezas y numerosos resultados interconectados (Quadri & Kalyankar, 2010).

Máquinas de soporte vectorial.

Las máquinas de soporte vectorial – (SVM) nace en los años 90, fundamentados en la teoría del aprendizaje estadístico, fueron planteadas por (Boser et al., 1992). Ideadas para resolver preocupaciones de la clasificación binaria, en la actualidad son empleadas para corregir

otras dificultades (agrupamiento, multiclasificación, regresión). De igual modo, son distintas las áreas en las que han sido usadas con agrado, así como identificación de imágenes, comprobación de caracteres, clasificación de texto e hipertexto, agrupación de proteínas, procesamiento de lenguaje natural, estudio de series temporales (Boser et al., 1992).

La máquina de soporte vectorial conceptualmente implementa la siguiente idea: los vectores de entrada son mapeados no linealmente a un ámbito de atributos de muy alta extensión. En este espacio de atributos se construye una extensión de decisión lineal (Cortes & Vapnik, 1995). Las propiedades especiales de la extensión de decisión aseguran una alta capacidad de generalización de la máquina de aprendizaje. La idea detrás de la red de apoyo vectorial se implementó anteriormente para el caso restringido en el que los datos de formación pueden separarse sin errores (Cortes & Vapnik, 1995).

En (Han et al., 2012) indican que un SMV utiliza un mapeo no lineal para transformar los datos originales de la formación en una dimensión superior. Dentro de esta nueva dimensión, busca el hiperplano de separación óptimo lineal (es decir, un "límite de decisión" que separa las tuplas de una clase de otra).

Los SVMs pueden ser usados para la predicción numérica, así como para la clasificación. Se han aplicado a una serie de áreas, incluyendo el reconocimiento manuscrito de dígitos, el reconocimiento de objetos y la identificación de hablantes, así como pruebas de predicción de series de tiempo de referencia (Han et al., 2012).

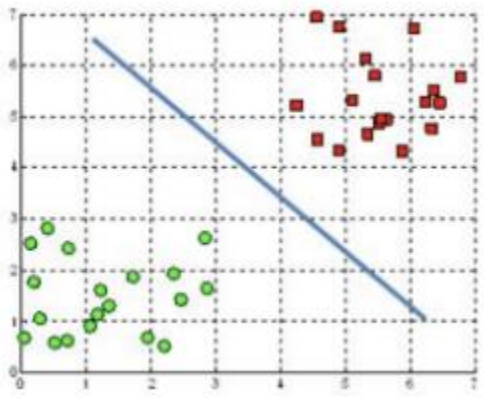


Figura 4. Hiperplano en el espacio de características 2D

Fuente: (Gandhi, 2018)

El objetivo del algoritmo de la máquina de soporte vectorial es encontrar un hiperplano en un espacio de dimensiones N (N - el número de características) que clasifique claramente los puntos de datos. Para separar las dos clases de puntos de datos, hay muchos hiperplanos posibles que podrían ser elegidos. El objetivo es encontrar un plano que tenga el margen máximo, es decir, la distancia máxima entre los puntos de datos de ambas clases. Maximizar la distancia del margen proporciona algún refuerzo para que los futuros puntos de datos puedan ser clasificados con más confianza (Carmona Suárez, 2014; Saravanan & Sujatha, 2018).

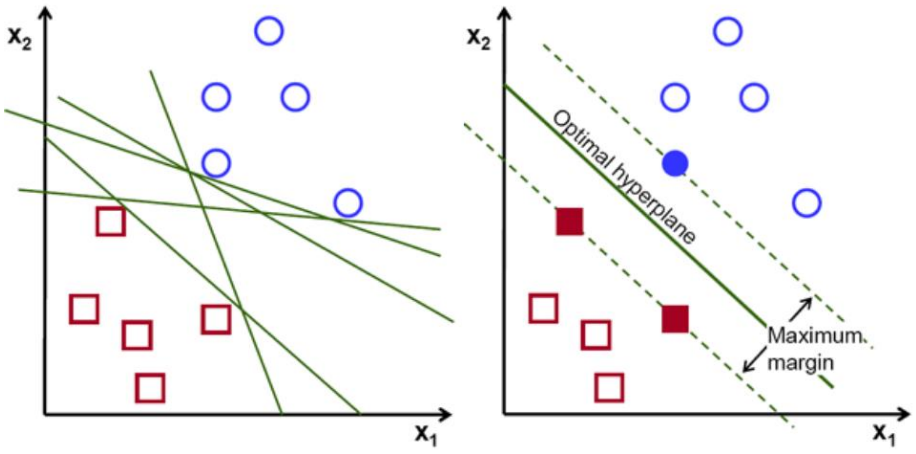


Figura 5. Ejemplos de hiperplanos

Fuente: (Gandhi, 2018)

Fortalezas de SVM

En (Betancourt, 2005) destacan las siguientes fortalezas de las máquinas de soporte vectorial:

- Entrenamiento fácil
- No existe un óptimo local
- Escalamiento bueno para datos en espacio con dimensiones altas
- La responsabilidad entre el error y la complejidad del clasificador puede ser contrastado claramente.
- Cadenas de caracteres y arboles pueden ser usados como datos no tradicionales como entrada a la SVM, en vez de vectores de características

Debilidades de SVM

Es necesario contar con una buena función kernel, en otras palabras, se requieren procedimientos eficientes para recibir los parámetros de inicialización de la SVM" (Betancourt, 2005).

Clasificadores Bayesianos.

La clasificación es una labor sencilla en el estudio de datos y la identificación de patrones que necesita la construcción de un clasificador, en otros términos, una función que concede una etiqueta de clase a las instancias descritas por un conjunto de atributos (Friedman et al., 1997). Las redes bayesianas se consideran una posibilidad a los sistemas expertos clásicos dirigidos a la toma de decisiones y al pronóstico bajo indecisión en lo referente a probabilísticos (Picard et al., 2004). Así mismo, en (García et al., 2006) plantean que las redes de Bayes son herramientas estadísticas orientadas a la modelación causal. Una red de Bayes tiene dos componentes principales: una dimensión cualitativa basada en la teoría de los grafos y una cuantitativa basada en la teoría de la probabilidad.

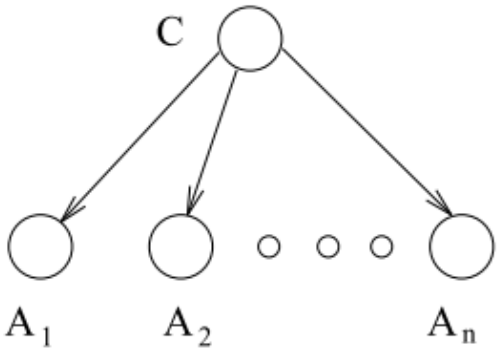


Figura 6. Estructura de la red Bayesiana

Fuente: (Friedman et al., 1997)

Una red bayesiana está compuesta por nodos de casualidad, unidos con otros nodos por arcos. Estos arcos están dirigidos: el arco va de un nodo padre a un nodo hijo. El estado en el que se encuentra un nodo padre no se ve afectado por el estado en el que se encuentran sus hijos. Pero la probabilidad de cada uno de los estados en los que puede estar un nodo hijo depende del producto cartesiano de los estados actuales de todos sus nodos padres (Edwards & Fasolo, 2001; Picard et al., 2004).

El clasificador de red Bayesiana aprende de los datos de preparación la probabilidad eventual de cada atributo A_i dada la etiqueta de clase *C*. La clasificación se hace entonces adaptando la regla de Bayes para evaluar la probabilidad de *C* dada la instancia especifica de A_1,...,A_n, y luego prediciendo la clase con la probabilidad posterior más alta (Friedman et al., 1997).

Los clasificadores tales como, Arboles de Decisión, Maquinas de Soporte Vectorial y Redes Bayesianas, son ampliamente utilizados para la predicción de datos de diferentes estudios (Cheewaprakobkit, 2013; Hasbun et al., 2016; Kalles & Pierrakeas, 2006b; Krishna Kishore et al., 2014; Lykourentzou et al., 2009; Márquez-Vera et al., 2013; Moseley & Mead, 2008; Veitch,

2004), efectuando comparativos que les permiten determinar cuál clasificador predice de mejor manera un conjunto de datos. Para eso son utilizadas las Métricas de Evaluación, tales como: Falsos Positivos, Falsos Negativos, Verdaderos Positivos, Falsos Positivos, Matriz de Confusión, entre otras, las cuales serán descritas a continuación.

Métricas de evaluación

Hay cuatro métricas básicas de calidad (Falsos Positivos, Falsos Negativos, Verdaderos Positivos y Verdaderos Negativos). Dichas métricas hacen parte la llamada Matriz de Confusión o Matriz de Clasificación (S. Kotsiantis et al., 2004). En la tabla 1 se explica de mejora forma la matriz de confusión.

"La matriz de confusión de clase n es una matriz nxn en la que las filas se nombran según las clases reales y las columnas, según las clases previstas por el modelo. Sirve para mostrar de forma explícita cuándo una clase es confundida con otra. Por eso, permite trabajar de forma separada con distintos tipos de error" (Recuero, 2018).

Tabla 1.

Matriz de Confusión

Matriz de Confusión		**Predicción**	
		Negativo	Positivo
Real	Negativo	a	b
	Positivo	c	d

Tomado de (Recuero, 2018)

- a: número de predicciones correctas de la clase negativos (falsos reales)
- b: número de predicciones incorrectas de la clase positiva (falsos positivos)
- c: número de predicciones incorrectas de la clase negativa (falsos negativos)

- d: número de predicciones correctas de la clase positiva (positivos reales)

A partir de las métricas básicas de la matriz de confusión se calculan otras métricas útiles, a continuación, se mencionará dichas métricas.

Precisión.

Precisión de clasificación es lo que normalmente se dice cuando se usa el término precisión. Es la conexión entre el número de predicciones correctas y el número total de muestras de entrada (A. Mishra, 2018).

$$Precisión = \frac{\text{Número de predicciones correctas}}{\text{Total de número de predicciones realizadas}} \quad (1)$$

Funciona bien sólo si hay igual número de muestras pertenecientes a cada clase. Por ejemplo, considere que hay 98% de muestras de clase A y 2% de clase B en nuestro set de entrenamiento. Entonces, nuestro modelo puede obtener fácilmente una precisión de entrenamiento del 98% con sólo predecir cada muestra de entrenamiento perteneciente a la clase A (A. Mishra, 2018).

Cuando se prueba el mismo modelo en un equipo de prueba con 60% de muestras de clase A y 40% de muestras de clase B, la precisión de la prueba se reduce al 60%. La precisión de la clasificación es grande, pero nos da la falsa sensación de lograr una alta precisión (A. Mishra, 2018).

El verdadero problema surge cuando el costo de la clasificación errónea de las muestras de las clases menores es muy alto. Si tratamos con una enfermedad rara pero fatal, el costo de no diagnosticar la enfermedad de una persona enferma es mucho más alto que el costo de enviar a una persona sana a más pruebas (A. Mishra, 2018).

Sensibilidad y Especificidad.

"La sensibilidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como positivo respecto a la condición que estudia la prueba, razón por la que también es denominada fracción de verdaderos positivos (FVP)" (López de Ullibarri & Píta Fernández, 1998). "La especificidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como negativo. Es igual al resultado de restar a uno la fracción de falsos positivos (FFP)" (López de Ullibarri & Píta Fernández, 1998).

Tabla 2.

Resultado de una prueba y su estado respecto a la enfermedad

		Verdadero diagnóstico	
		Enfermo	**Sano**
Resultado de la Prueba	Prueba Positiva	Verdadero Positivo (VP)	Falso Positivo (FP)
	Prueba Negativa	Falso Negativo (FN)	Verdadero Negativo (VN)
		VP + FN	**VN + FP**

Tomado de (López de Ullibarri & Píta Fernández, 1998)

Basado en la Tabla 2, el cálculo de la sensibilidad y especificidad se calcula de la siguiente manera:

- Sensibilidad: VP / (VP + FN) = FVP (fracción de verdaderos positivos)
- Especificidad: VN / (VN + FP) = FVN (fracción de verdaderos negativos)

Curvas ROC.

Las curvas ROC se progresaron en los años cincuenta como instrumentos para el estudio de descubrimiento y análisis de señales de radar (Burgueño et al., 1995). Los autores definen que "la curva ROC es un gráfico en el que se observan todos los pares sensibilidad/especificidad resultantes de la variación continua de los puntos de corte en todo el rango de resultados observados". En el eje y se ubica la sensibilidad o fracción de verdaderos positivos.

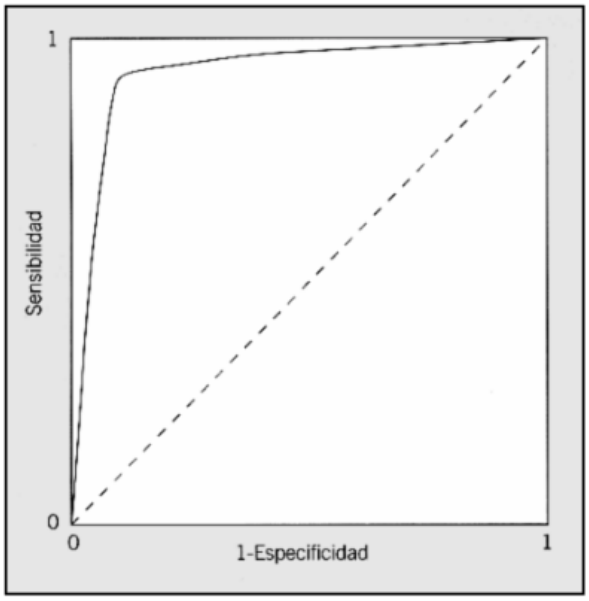


Figura 7. Curva ROC

Tomada de (Burgueño et al., 1995)

Las curvas ROC son indicadores de la precisión señalada y facilitan un principio unitario en el proceso de valoración de un examen, a causa de sus distintas aplicaciones (Burgueño et al., 1995). Los autores definen las siguientes ventajas del uso de curva ROC:

- Representación fácil y clara de la suficiencia de distinción de la prueba en toda la clase de puntos de corte.
- Gráficas sencillas y simples de explicar visualmente

- No necesitan un nivel de resolución específico, debido a que está incorporado toda la visión de puntos de corte.
- Es autónoma de la prevalencia, debido a que la sensibilidad y la especificidad se consiguen en diferentes subgrupos. Por tanto, no es necesario tener cuidado para obtener muestras con prevalencia característica de la población. En realidad, es mejor regularmente disponer de idénticos números de individuos en ambos subgrupos.
- Suministran un contraste visual claro entre pruebas en una proporción general, por el contrario, otra clase de gráficos, como los histogramas de frecuencias o los diagramas de puntos, necesitan distintos gráficos cuando discrepan las escalas.
- "La especificidad y la sensibilidad son accesibles en el gráfico, en contraste con los diagramas de puntos y los histogramas" (Burgueño et al., 1995).

Media Armónica.

"La media armónica se define como el recíproco de la media aritmética de los recíprocos" (Universidad Pedagógica y Tecnológica de Colombia, 2004) :

$$MA = \frac{1}{\frac{1}{n}(\frac{1}{X_1}+\frac{1}{X_2}+\cdots+\frac{1}{X_n})} \qquad (2)$$

El uso de la media geométrica o de la armónica corresponde a una modificación de la variable en log *X* o 1/*X* (Universidad Pedagógica y Tecnológica de Colombia, 2004).

Cobertura o Recall.

En (Han et al., 2012) definen la Cobertura como una medida de integridad qué indica el porcentaje de tuplas positivas se etiquetan como tales. Es lo mismo que la sensibilidad (o la verdadera tasa positiva) (Han et al., 2012). Se calcula de la siguiente manera:

$$\text{Recall} = \frac{TP}{TP+FP} \qquad (3)$$

Verdaderos positivos (TP): Estos se refieren a las tuplas positivas que fueron correctamente etiquetadas por el clasificador. Dejemos que TP sea el número de verdaderos positivos.

Negativos verdaderos (TN): Estas son las tuplas negativas que fueron correctamente etiquetadas por el clasificador.

Falsos positivos (FP): Estas son las tuplas negativas que fueron etiquetadas incorrectamente como positivas.

Falsos negativos (FN): Estas son las tuplas positivas que fueron mal etiquetadas como negativas.

Investigaciones relacionadas

Existe un amplio espectro de investigaciones relacionadas a la planteada en este proyecto. A continuación, se mencionan las categorías identificadas en los casos de estudio. Análisis de Satisfacción: (Thomas & Galambos, 2004). Predicción del rendimiento estudiantil: (Aziz et al., 2015; S. Kotsiantis et al., 2004). Riesgo de reprobación: (Kalles & Pierrakeas, 2006a; Kalles & Pierrakeas, 2006b; Barnes et al., 2009; Delen, 2010; Zhang & Oussena, 2010; Tsai et al., 2011; Cheewaprakobkit, 2013; Manhães et al., 2014; Sharabiani et al., 2014; Sangodiah et al., 2015).

Análisis de retención: (Salazar et al., 2004; Veitch, 2004; Moseley & Mead, 2008; Lykourentzou et al., 2009; Delen, 2011; Jin et al., 2011; Alkhasawneh & Hobson, 2011; Bayer et al., 2012; Mustafa et al., 2012; Márquez-Vera et al., 2013; Pereira et al., 2013; Sarker et al., 2014; T. Mishra et al., 2014; Barbosa Manhães et al., 2014; Krishna Kishore et al., 2014; Güner et al., 2014; Siri, 2015; Chai & Gibson, 2015; Şara et al., 2015; Cambruzzi et al., 2015; Santana et al., 2015; Lesinski et al., 2016; Devasia et al., 2016; Hernandez Gonzalez et al., 2016; Hasbun et al., 2016; Askinadze & Conrad, 2017; Zeng et al., 2017; Miranda & Guzmán, 2017; Dharmawan et al., 2018; Segura-Morales & Loza-Aguirre, 2018; Solis et al., 2018; Pérez et al., 2019; Beaulac & Rosenthal, 2019;Lee & Chung, 2019). Cada una de estas categorías con los estudios serán detallados a continuación:

Analisis de satisfacción

En (Thomas & Galambos, 2004) realizan el estudio en las características y experiencias que afectan la satisfacción de los estudiantes. Investigando en medidas alternativas de la satisfacción general de los estudiantes utilizando regresión múltiple y análisis de árbol de decisión con el algoritmo del detector automático de interacción chi-cuadrado (CHAID). Los datos para este análisis provienen de una encuesta de opinión de estudiantes en una universidad pública de investigación. La encuesta de opinión de los estudiantes recopila datos sobre una amplia gama de características, experiencias y planes de los estudiantes; su satisfacción con el ambiente, clima, servicios e instalaciones del campus; sus percepciones del crecimiento; y las razones de su elección de universidad. Del análisis realizado por los algoritmos, hay un 93% de estudiantes que reportan un crecimiento intelectual muy grande. El autor indica que no sólo los factores académicos, sino también las variables sociales y de servicio son predictores de la satisfacción de estos estudiantes con la calidad de la educación.

Predicción del rendimiento estudiantil

En el siguiente estudio los autores (S. Kotsiantis et al., 2004) se enfocaron en la identificación de estudiantes con bajo rendimiento académico en un sistema de aprendizaje a distancia. Los autores utilizaron un conjunto de datos proporcionado por la Universidad Abierta Helénica en Grecia, del curso de informática de la misma Universidad. Tomando los registros de 510 estudiantes matriculados en el periodo 2000-1. Analizaron variables socioeconómicas como la edad, sexo, estado civil, ocupación, número de hijos, conocimientos de informática, trabajo asociado a ordenadores y atributos académicos. Aplicaron 6 técnicas de minería, arrojando que el mejor algoritmo fue Naıve Bayes (72,48%), seguido por el de Regresión Logística (72,32%), el de BP (Back Propagation) (72,26%) y el de SMO (72,17%). No hubo diferencias estadísticamente significativas entre los cuatro anteriores. Por el contrario, el algoritmo C4.5 (69.99%) que sigue es de menor precisión estadísticamente significativa que todos los algoritmos mencionados anteriormente ($p < 0.001$). Los autores (S. B. Kotsiantis & Pintelas, 2005) utilizando el conjunto de datos anterior y aplicando técnicas de minería de datos y resultados obtenidos, crearon un software prototipo para clasificar las dificultades de aprendizaje de los estudiantes, permitiendo la toma de decisiones y determinando las dificultades de los estudiantes de la Universidad Abierta Helénica.

Los autores (Aziz et al., 2015) proponen un framework predecir el rendimiento de los estudiantes de primer año de licenciatura en Ciencias de la Computación. El clasificador Naïve Bayes se usó para extraer patrones utilizando WEKA como herramienta de minería de datos con el fin de construir un modelo de predicción. Los datos se recopilaron a partir de registros de seis años comprendido entre julio de 2006/2007 y julio de 2011/2012. De los datos de los estudiantes, se seleccionaron seis parámetros que son raza, género, ingreso familiar, modo de ingreso a la

universidad y promedio de calificaciones. Usando el clasificador Naïve Bayes, predicirá la etiqueta de la clase "Grade Point Average" como un valor categórico; Pobre (Poor), Promedio (Average), y Bueno (Good). El resultado del estudio muestra que los ingresos familiares de los estudiantes, el género y los parámetros de la ciudad natal contribuyen al rendimiento académico de los estudiantes.

Riesgo de reprobación

En (Kalles & Pierrakeas, 2006a) tomaron como referencia las investigaciones realizadas por (S. Kotsiantis et al., 2004; S. B. Kotsiantis & Pintelas, 2005) en la Universidad Abierta Helénica, confirmando su validez desde un punto de vista cualitativo y produciendo modelos consistentemente más cortos, con una configuración básica de los parámetros para la derivación de modelos, basados en una combinación de enfoques de aprendizaje automático, es decir, árboles de decisión y algoritmos genéticos. Utilizaron WEKA para analizar el conjunto de datos, la técnica que arrojó mejor accuracy (92.50%) fueron los árboles de decisiones (J48).

Los autores (Kalles & Pierrakeas, 2006b) usaron algoritmos genéticos para derivar árboles de decisión cortos que explican el fracaso de los estudiantes. Dichos fracasos son reportados en los cursos introductorios en la Universidad Abierta Helénica, la cual ofrece carreras a distancia, con más de 25.000 estudiantes inscritos para el 2005. Los autores indican que en el aprendizaje abierto y a distancia, las tasas de deserción estudiantil más altas que las de las universidades convencionales. Usando los conjuntos de datos de los estudiantes para desarrollar modelos de éxito/fracaso representados como árboles de decisión. Los resultados obtenidos en las experimentaciones realizadas arrojaron una precisión entre el 78% y 84.61 para predicción de las tasas de fracaso de los estudiantes, tomando mejores decisiones en los tutores y estrategias que ayuden a mitigar la deserción y fracaso de los cursos tomados por los estudiantes.

En la 2ª Conferencia Internacional de Minería de Datos Educativos (Barnes et al., 2009), los autores (*Predicting Students Drop Out A Case Study*, 2009) plantearon un caso de estudio donde describieron los resultados del estudio de caso de minería de datos educativos (EDM) destinado a predecir el abandono de los estudiantes de Ingeniería Eléctrica (EE) después del primer semestre de sus estudios o incluso antes de que ingresen al programa de estudios, así como a identificar los factores de éxito específicos del programa académico. Los resultados experimentales mostraron que clasificadores bastante simples e intuitivos (árboles de decisión) dan un resultado útil con precisiones entre el 75 y el 80%. Además, demostraron la utilidad de un aprendizaje sensible a los costes y un análisis exhaustivo de las clasificaciones erradas, mostrando algunas formas de mejorar la predicción sin tener que recopilar datos adicionales sobre los estudiantes.

En (Delen, 2010) utilizaron cinco años de datos institucionales junto con varias técnicas de minería de datos (tanto individuos como conjuntos), el autor desarrolló modelos analíticos para predecir y explicar las razones del desgaste de los estudiantes de primer año. Los resultados de los análisis comparativos mostraron que los conjuntos funcionaron mejor que los modelos individuales, mientras que el conjunto de datos equilibrado produjo mejores resultados de predicción que el conjunto de datos desequilibrado. El análisis de sensibilidad de los modelos reveló que las variables educativas y financieras se encuentran entre los predictores más importantes del fenómeno. las máquinas de soporte vectorial produjeron los mejores resultados con una tasa de predicción global del 87,23%, y el árbol de decisión quedó en segundo lugar con una tasa de predicción global del 87,16%, seguido de las redes neuronales artificiales y la regresión logística con tasas de predicción global del 86,45% y el 86,12% respectivamente.

En (Zhang & Oussena, 2010) los autores argumentan cómo la minería de datos puede ayudar a detectar a los estudiantes "en riesgo", evaluar la idoneidad del curso o módulo y adaptar las intervenciones para aumentar la retención de estudiantes. Para ello, utilizaron información social de los estudiantes e información académica. Los objetivos se centraban en: Detectar patrones de Comportamiento del Estudiante, Detectar patrones de Comportamiento del Curso, Predecir de la Retención del Estudiante, Predecir de la Idoneidad del Curso, Crear una Estrategia de Intervención Personalizada. Naive Bayes obtuvo una predicción de 85.9%, las máquinas de soporte vectorial obtuvieron 78.7% y en último lugar los árboles de decisiones con 71.2%.

Por otro lado, los autores (Tsai et al., 2011) usaron técnicas de minería de datos para hacer predicciones acerca de los estudiantes que van a tomar el examen de competencia en computación y fallar. Una universidad nacional en Taiwán se usó como el caso de estudio. Los autores utilizaron tres técnicas de agrupación diferentes para agrupar a los estudiantes en diferentes grupos, que son k-means, mapas auto-organizados (SOM) y agrupación en dos pasos (BIRCH). Después de encontrar el mejor resultado de clusterización, se utiliza el algoritmo del árbol de decisión para extraer reglas útiles de cada uno de los clusters identificados. Estas reglas se pueden usar para advertir o aconsejar a los estudiantes que tienen mayor probabilidad de reprobar el examen. Después de identificar los mejores resultados de clustering y sus características, el siguiente paso fue extraer las reglas de decisión de cada uno de los clusters. Los conjuntos de entrenamiento y pruebas para construir y probar el modelo de árbol de decisión se basan en BIRCH (k = 5). Al usar las reglas de decisión se pudo predecir correctamente que alrededor del 80% de los estudiantes que no pasarán la prueba de competencia en computación en el conjunto de pruebas.

En (Cheewaprakobkit, 2013) analizan los factores que afectan el logro académico que contribuye a la predicción del desempeño académico de los estudiantes. Útil para identificar a los estudiantes débiles que tienen probabilidades de tener un desempeño deficiente en sus estudios. El conjunto de datos comprendía 1.600 registros de estudiantes con 22 atributos de estudiantes matriculados entre el año 2001 y 2011 en una universidad de Tailandia. El investigador aplicó el conjunto de datos para diferenciar los clasificadores (árbol de decisión, red neuronal). Se utilizó una validación cruzada con 10 pliegues para evaluar la precisión de la predicción. Los resultados muestran que el clasificador del árbol de decisión alcanza una alta precisión del 85,188%, que es superior en un 1,313% a la del clasificador de la red neuronal.

Los autores (Manhães et al., 2014) usaron técnicas de minería de datos educativos (EDM) para identificar las variables que pueden ayudar a los gestores educativos a detectar a los estudiantes que presentan un bajo rendimiento o que están en riesgo de abandonar sus estudios de pregrado. Usando datos académicos de estudiantes de la mayor Universidad Pública Federal Brasileña. Establecieron tres categorías de estudiantes con diferente trayectoria académica para investigar su desempeño y las tasas de deserción escolar. El estudio mostró que incluso analizando tres clases diferentes de 14.000 estudiantes fue posible tener una precisión global superior al 80% para varios algoritmos de clasificación. Se utilizaron los resultados del modelo de Naïve Bayes para apoyar el análisis cuantitativo.

Los autores (Sharabiani et al., 2014) presentaron un modelo para predecir el rendimiento académico de los estudiantes de los estudiantes de Ingeniería. El modelo se basa en el marco de las redes bayesianas. El modelo lo construyeron utilizando una base de datos de los estudiantes de ingeniería de la Universidad de Illinois en Chicago (UIC). El objetivo específico de este

modelo es predecir las calificaciones de los estudiantes en tres cursos principales que la mayoría de los estudiantes toman en su segundo semestre. Logrando predicciones mayores a 73%.

En el siguiente estudio (Sangodiah et al., 2015) se centraron en el uso del modelo de máquinas de soporte vectorial para predecir el estado de libertad condicional del estudiante, en el que en la mayoría de los casos conducirá a la expulsión del estudiante. También se examinarán los factores pertinentes y de otro tipo que contribuyen al desgaste entre los estudiantes de Malasia. Los autores concluyeron que los resultados del estudio fueron interesantes, ya que el modelo de logró una precisión decente en la predicción a pesar de trabajar con un conjunto de datos de pequeño tamaño, 89.84%.

Análisis de retención

En (Salazar et al., 2004) presentaron un estudio de investigación aplicada sobre el descubrimiento de conocimientos basado en el análisis de datos académicos. Los objetivos principales de este estudio de investigación fueron obtener conocimiento sobre el éxito y el fracaso académico, la retención y la deserción estudiantil. Se utilizaron técnicas de minería de datos de clustering automático y reglas de decisión. La aplicación del algoritmo C-mean a subconjuntos de datos estadísticamente homogéneos proporcionó un grupo de conglomerados, que se han descrito cualitativamente. Utilizando clusters seleccionados, un estudio de reglas de decisión basado en el algoritmo C4.5 generó un conjunto de reglas de decisión para los cuatro temas de investigación del estudio. La información del estudio fue basada de datos académicos de la Universidad Industrial de Santander (IUS). Como la facultad, el programa académico, el género, la categoría de estudiante, el área de origen, etc. También contiene los datos de los estudiantes datos sobre el rendimiento académico, como la nota media académica acumulada y la nota de la prueba nacional preuniversitaria.

En (Veitch, 2004) se detallan las correlaciones de la deserción escolar mediante el uso de la minería de datos de las fuentes de datos existentes con los árboles de decisión. Todas las variables utilizadas en el estudio se extrajeron directamente de las bases de datos electrónicas de los distritos. Los estudiantes de secundaria registrados como "abandonados" (sin registro de transferencia) en el curso del año académico 2001-2002 fueron emparejados con una muestra aleatoria de no abandonados. Utilizando los árboles de decisión, la predicción debe ser correcta más del 80% del tiempo.

En (Moseley & Mead, 2008) utilizaron arboles de decisiones para predecir la deserción de estudiantes de enfermería de una Universidad Británica. Empleando información socioeconómica y académica como: edad, sexo, notas del estudiante, promedio de los semestres, entre otros. Recopilando datos de 528 estudiantes en 5 años. Utilizando 3978 registros únicos, divididos en un conjunto de entrenamiento y un conjunto de pruebas. Obtuvieron una sensibilidad del 84%, especificidad del 70% y una precisión del 94%. Los autores argumentan la necesidad de tener grandes cantidades de datos y de alta calidad.

Los autores (Lykourentzou et al., 2009) presentan un método de predicción de deserción educativa para identificar con precisión a los estudiantes propensos a la deserción durante las primeras etapas del curso de aprendizaje electrónico. El método propuesto aprovecha las características de los cursos de e-learning, que incluyen datos detallados de actividad y progreso de los estudiantes, para realizar sus predicciones de forma dinámica y adaptarlas en función de los niveles de rendimiento y participación de cada alumno a lo largo del curso. Estos datos se utilizan para entrenar tres técnicas de aprendizaje automático: las redes neuronales de retroalimentación (FFNN), las máquinas de soporte vectorial (SVM) y el conjunto probabilístico de ARTMAP difuso simplificado (PESFAM). Utilizando estos esquemas, el método propuesto

logró una tasa de clasificación general de estudiantes del 75-85%. Los resultados relativos a los criterios de sensibilidad y precisión también fueron elevados, lo que indica que el esquema era preciso tanto en la identificación correcta de los abandonos como en la prevención de errores de clasificación.

En este estudio (Delen, 2011), usando 8 años de datos institucionales junto con tres técnicas populares de minería de datos, el autor desarrolló modelos analíticos para predecir la deserción de los estudiantes de primer año. De los tres tipos de modelos (redes neuronales artificiales, árboles de decisión y regresión logística), las redes neuronales artificiales se desempeñaron mejor, con una precisión de predicción general del 81% en la muestra de retención. El análisis de importancia variable de los modelos reveló que las variables educativas y financieras son las más importantes entre los predictores utilizados en este estudio.

Los autores (Jin et al., 2011) proponen un modelo de red neuronal de retropropagación para predecir la retención y el GPA (Grade Point Average) universitario de los estudiantes de ingeniería. Usando datos de 1470 estudiantes de ingeniería de primer año que se matricularon en una gran universidad del Medio Oeste durante el año académico 2004-2005. la precisión general de la predicción de retención es del 71,3%.

En (Alkhasawneh & Hobson, 2011) desarrollaron dos modelos de redes neuronales que utilizan una red de retropropagación de retroalimentación para predecir la retención de los estudiantes en los campos de la ciencia y la ingeniería. El primer modelo se utiliza para predecir la retención de los estudiantes de primer año e identificar los factores preuniversitarios correlacionados. El segundo modelo es clasificar a los grupos de estudiantes de primer año en tres clases: estudiantes en riesgo, intermedios y avanzados. Con un total de 338 muestras utilizadas, el 70,1% de los estudiantes clasificaron correctamente.

En (Bayer et al., 2012) se centran en predecir el abandono escolar y el fracaso escolar cuando los datos de los estudiantes se han enriquecido con datos derivados del comportamiento social de los estudiantes. Estos datos describen las dependencias sociales recopiladas a partir del correo electrónico y de conversaciones en foros de discusión, entre otras fuentes. El conjunto de datos contenía 775 estudiantes, 837 estudios y 4.373 ejemplos en total. Los autores detallan extracción de nuevas características tanto de los datos de los estudiantes como de los datos de comportamiento representados por un gráfico social construido. Luego presentaron un método novedoso para el aprendizaje de un clasificador para la predicción del fracaso de los estudiantes que emplea el aprendizaje sensible a los costos para reducir el número de estudiantes clasificados incorrectamente como no exitosos. Implementaron arboles de decisiones, máquinas de soporte vectorial y Naive Bayes. El algoritmo que logró la mejor tasa de precisión fue el árbol de decisión PART, con una precisión del 93.67% y una tasa de true positive (TP) de 92.30%. Demostraron que el uso de datos de comportamiento social resulta en un aumento significativo de la precisión de la predicción.

El objetivo principal de (Mustafa et al., 2012) fue desarrollar un modelo dinámico de predicción de deserción escolar para universidades, institutos y colegios. Aplicando la prueba de chi cuadrado a factores separados tales como el género, la condición financiera y el año de caída para clasificar a los estudiantes exitosos de los que no lo son. El propósito principal de aplicarla es la selección de características a los datos. El grado de libertad se utiliza para calcular el valor P (valor de probabilidad) para los mejores predictores de la variable dependiente. Después de la separación de factores los autores examinaron mediante técnicas de minería de datos de Clasificación y Árbol de Regresión (CART) y Árbol de CHAID. Entre los árboles de clasificación, los métodos de crecimiento Clasificación y Árbol de Regresión (CART) fueron los

más exitosos en el crecimiento del árbol con un porcentaje general de clasificación correcta que el árbol CHAID.

En (Márquez-Vera et al., 2013) los autores aplicaron técnicas de minería de datos para predecir el fracaso escolar y la deserción escolar. Utilizaron datos reales de 670 estudiantes de la Universidad Autónoma de Zacatecas (UAPUAZ) para el año académico 2009/10, y empleando métodos de clasificación, tales como reglas de inducción y árboles de decisión. Los experimentos intentan mejorar su precisión para predecir qué estudiantes podrían fracasar o abandonar los estudios, primero, utilizando todos los atributos disponibles (77); luego, seleccionando los mejores atributos (15); y finalmente, reequilibrando los datos y utilizando una clasificación sensible a los costos. La precisión más alta la obtuvieron los árboles de decisión con una tasa de aciertos de 96.6%.

Así mismo, en (Pereira et al., 2013) presentan los de un proyecto de investigación que busca identificar patrones de deserción escolar a partir de datos socioeconómicos, académicos, disciplinarios e institucionales de estudiantes de pregrado de la Universidad de Nariño de la ciudad de Pasto (Colombia), utilizando técnicas de minería de datos. Crearon un conjunto de datos con los registros de los estudiantes que fueron admitidos en los períodos comprendidos entre el primer semestre de 2004 y el segundo semestre de 2006. Se analizaron tres cohortes completas con un período de observación de seis años hasta 2011. Se descubrieron los perfiles socioeconómicos y académicos de los estudiantes que abandonaron la escuela utilizando técnicas de clasificación basadas en árboles de decisión. Logrando una precisión superior a 80%.

En el siguiente estudio los autores (Sarker et al., 2014) detallan un modelo de predicción de estudiantes que utiliza datos abiertos externos disponibles comúnmente en lugar de los cuestionarios/encuestas tradicionales para detectar los estudiantes en riesgo. Para crear el

conjunto datos, los autores utilizaron información académica, información de cuestionarios, e información de datos externos. El número total de participantes en este estudio fue de 149, de los cuales cerca del 15% están en el grupo de estudiantes "en riesgo" y el 85% en el grupo de estudiantes "no en riesgo". Los autores usaron en este estudio perceptrón multicapa (MLP – MultiLayer Perceptron), la precisión del mejor de los tres modelos presentados por los autores fue de 90.44% con una sensibilidad del modelo de 73% y una especificidad de 93%.

En (T. Mishra et al., 2014) utilizaron diferentes técnicas de clasificación para construir un modelo de predicción del rendimiento basado en la integración social de los estudiantes, la integración académica y varias habilidades emocionales. Dos algoritmos J48 (Implementación de C4.5) y Random Tree han sido aplicados a los registros de los estudiantes de MCA de las universidades afiliadas a Guru Gobind Singh Indraprastha University, en delhi India, para predecir el desempeño del tercer semestre. Random Tree es más preciso para predecir el rendimiento que el algoritmo J48, logrando una precisión de 94.418%, versus 88.372% obtenido por J48.

Basados en el estudio anterior, los autores (Barbosa Manhães et al., 2014) diseñaron una arquitectura que utiliza técnicas de EDM (Educational Data Mining) para predecir e identificar a aquellos que están en riesgo de abandono escolar. Este enfoque permite a los gerentes académicos monitorear el progreso de los estudiantes en cada semestre académico, identificando a los que se encuentran en dificultades para cumplir con sus requisitos académicos. todos los clasificadores, están por encima del 87%. Las tasas negativas reales (aprobadas) son superiores a 0,9 para todos los clasificadores. Sin embargo, el objetivo principal es identificar al clasificador con una tasa más alta para el verdadero positivo (no aprobado). El clasificador Naïve Bayes

presentó la tasa realmente positiva más alta para todos los conjuntos de datos utilizados en los experimentos.

Los autores (Krishna Kishore et al., 2014) proponen una aplicación de predicción basada en la Percepción Multicapa (MLP) para predecir el Promedio de Calificaciones (GPA) de los estudiantes de pregrado mediante el uso de la Historia Académica Previa del estudiante, la Regularidad, el Número de Atrasos, el Grado de Inteligencia, la Naturaleza del Trabajo, la Disciplina, las Actividades Sociales y el Grado. Con esta aplicación fue posible predecir los datos del estudiante que están en riesgo, y algunas medidas proactivas como clases extra y material de apoyo se ofrecen para mejorar el progreso académico de esos estudiantes. Los autores utilizaron datos de 134 estudiantes de tercer año de Ingeniería Informática de la Universidad de Vignan en India, logrando una precisión de predicción del 97,37% con MLP.

En (Güner et al., 2014) presentan un estudio sobre la predicción de los estudiantes de ingeniería en situación de riesgo académico en las primeras etapas. Utilizando máquinas de soporte vectorial SVM y redes neuronales artificiales ANN. La población de estudio incluyó a todos los estudiantes matriculados en la Facultad de Ingeniería de la Universidad de Pamukkale en Turquía en los periodos académicos 2008-2009 y 2009-2010 con estudiantes de primer año. Los datos se obtuvieron de varias instituciones y se realizan cuestionarios a los estudiantes. Cada punto de entrada de datos es de 38 características, que incluye información demográfica y académica sobre los estudiantes, mientras que la salida basada en el promedio de notas del primer año de los estudiantes cae en riesgo o no. Los resultados del estudio han demostrado que los métodos de máquinas vectoriales de apoyo o de redes neuronales artificiales se pueden utilizar para predecir el rendimiento de un estudiante en el primer año de forma prioritaria.

En (Siri, 2015) el autor investiga la primera etapa del proceso de transición del estudiante a la universidad. Con la ayuda de técnicas de minería de datos, en particular de las redes neuronales artificiales. El autor analizó registros de 810 estudiantes matriculados por primera vez en un curso de licenciatura en profesiones de la salud en la Universidad de Génova, Italia en el año académico 2008-09. La investigación se basó en el análisis de datos e información procedentes de fuentes primarias: datos administrativos relacionados con las carreras de los estudiantes; datos estadísticos recogidos durante la investigación mediante una encuesta; datos derivados de entrevistas telefónicas con estudiantes que no habían completado la matrícula en los años siguientes. La red neural predijo correctamente el 84 por ciento de los casos pertenecientes al grupo de desertores.

En (Chai & Gibson, 2015) desarrollaron un modelo de deserción estudiantil para predecir qué estudiantes de primer año están en mayor riesgo de abandonar la escuela en varios momentos durante su primer semestre. El objetivo de desarrollar un modelo de este tipo es ayudar a las universidades apoyando y reteniendo proactivamente a estos estudiantes a medida que su situación y riesgo cambian con el tiempo. El estudio evaluó diferentes modelos para predecir el desgaste de los estudiantes en cuatro períodos de tiempo diferentes a lo largo de un período de estudio semestral: modelos de preinscripción, matriculación, matriculación, durante el semestre y al final del mismo. Un conjunto de datos de 23.291 estudiantes que se matricularon en su primer semestre entre 2011-2013 fue extraído de varias fuentes de datos. Utilizaron regresión logística, árboles de decisión y bosques aleatorios. El rendimiento de estos modelos se evaluó utilizando las métricas de precisión y recuperación. El modelo logró el mejor rendimiento y utilidad para el usuario utilizando la regresión logística (67% de precisión, 29% de recuperación).

En (Şara et al., 2015) realizaron un estudio para predecir los estudiantes que no terminaran su educación secundaria. Los autores utilizaron la información del sistema de administración del estudio MaCom Lectio, que es utilizado por la mayoría de las escuelas secundarias danesas, con datos de fuentes públicas en línea (base de datos de nombres, planificador de viajes, estadísticas gubernamentales). Utilizaron una muestra de 36299 estudiantes para el conjunto de datos y 36299 para realizar pruebas. El clasificador que mejor obtuvo resultado fue Random Forrest, con una precisión del 93.47% y un área bajo la curva ROC por debajo de 0.965.

Por otro lado, en (Cambruzzi et al., 2015) presentan un sistema de Analítica de Aprendizaje desarrollado para tratar el problema de la deserción estudiantil en los cursos de educación a distancia en la universidad. Los autores emplearon varias herramientas complementarias, que permiten la visualización de datos, predicciones de deserción estudiantil, apoyo a acciones pedagógicas y análisis textuales, entre otras, están disponibles en el sistema. La implementación de estas herramientas es factible debido a la adopción de un enfoque llamado Multitrail para representar y manipular datos de varias fuentes y formatos. Los resultados obtenidos de los experimentos realizados con cursos en una universidad brasileña (do Vale do Rio dos Sinos - UNISINOS) muestran la predicción de deserción con un promedio de 87% de precisión, utilizando redes neuronales artificiales (ANN). Se implementó un conjunto de acciones pedagógicas relativas a los estudiantes con mayores probabilidades de deserción escolar y se observó una reducción promedio del 11% en las tasas de deserción escolar.

En (Santana et al., 2015) proponen modelos de predicción para proporcionar a los gestores educativos el deber de identificar a los estudiantes que se encuentran en el límite de deserción estudiantil. Utilizaron cuatro algoritmos de clasificación con diferentes métodos, con

el fin de encontrar el modelo con la mayor precisión en la predicción del perfil de los estudiantes que abandonaron los estudios. Los datos para la generación de modelos se obtuvieron de dos fuentes de datos disponibles en la Universidad Federal de Alagoas en Brasil. Los resultados mostraron que el modelo generado por el uso del algoritmo SVM es el más preciso entre los seleccionados, con una precisión del 92,03%.

En (Lesinski et al., 2016) presentan un enfoque de red neuronal para clasificar el estado de graduación de los estudiantes basado en indicadores académicos, demográficos y de otro tipo seleccionados. El modelo es entrenado, probado y validado usando 5100 muestras de estudiantes con datos recopilados de registros de admisión y bases de datos de investigación institucional. Nueve variables de entrada consisten en elementos de datos categóricos y numéricos que incluyen: rango en la escuela secundaria, calidad de la escuela secundaria, puntajes de exámenes estandarizados, evaluaciones del profesorado de la escuela secundaria, puntaje de actividad extracurricular, estado de educación de los padres y tiempo transcurrido desde la graduación de la escuela secundaria. El modelo fue capaz de predecir el éxito de la graduación y logró el mejor rendimiento con una precisión superior al 95%.

En (Devasia et al., 2016) desarrollaron una aplicación web que utiliza la técnica minera de datos naives bayes para la extracción de información útil. El experimento se llevó a cabo en 700 estudiantes con 19 atributos en Amrita Vishwa Vidyapeetham, Mysuru - India. El resultado demuestró que el algoritmo naive bayes proporciona más precisión que otros métodos como la regresión, el árbol de decisión y las redes neuronales, para la comparación y la predicción.

Por otra parte, en (Hernandez Gonzalez et al., 2016) presentan un estudio comparativo de predicción del riesgo de deserción escolar en el ITSM-México (Instituto Tecnológico Superior de Misantla-México). Este sistema utiliza la información personal y académica de los estudiantes

del ITSM. El estudio comparativo utiliza cuatro algoritmos: regresión logística, clustering, árboles de decisión y red neuronal, que consideran la información de la base de datos del sistema escolar de control del instituto. Los resultados muestran que el algoritmo de regresión logística tiene un buen acuerdo con los resultados experimentales.

En (Hasbun et al., 2016) estudia la importancia de las actividades extracurriculares para predecir la deserción escolar en estudiantes de dos licenciaturas (Ingeniería y Empresariales). Se recopilaron datos de 4.840 estudiantes y se capacitaron y validaron dos modelos, uno que incluye todos los datos y otro que elimina los créditos que valen la pena, lo que demuestra que las actividades extracurriculares son excelentes predictores de deserción escolar. La predicción lograda con los árboles de decisiones fue de 79.29%.

En (Askinadze & Conrad, 2017) examinaron la distancia de distorsión dinámica del tiempo (DTW) junto con el clasificador k-nn y mostraron cómo se puede utilizar el DTW como un núcleo SVM para la predicción de caídas en los datos de las series temporales. Con este enfoque, se reconocen alrededor del 67% de los abandonos del curso de estudios después del primer semestre y alrededor del 60% después del segundo semestre.

En (Zeng et al., 2017) demostraron resultados preliminares para predecir la deserción estudiantil de los estudiantes de cuidados en el hogar a partir de un amplio y heterogéneo conjunto de datos que contiene datos demográficos de los estudiantes y características de ingeniería extraídos de los patrones de entrenamiento. La predicción de la deserción estudiantil en diferentes etapas de capacitación de un estudiante adulto arrojó resultados a partir de un conjunto de datos sesgados de más de 5.303 estudiantes, siendo los árboles de decisiones el que arrojó las predicciones más sólidas, del 73%.

En (Miranda & Guzmán, 2017) usando datos entregados por las carreras de Ingeniería de la Universidad Católica del Norte en Antofagasta y Coquimbo (Chile), para determinar la importancia de las variables que conllevan a un estudiante a abandonar la Universidad. Usando técnicas de minería de datos, indican que la retención se sitúa en un 78%.

En (Dharmawan et al., 2018) tomaron datos de los estudiantes que estudian en varias universidades de Indonesia utilizando un muestreo aleatorio simple. Tomando información demográfica, motivación, financiera, interacción social y personalidad, analizaron con los clasificadores SVM, arboles de decisión y K-NN y obtener una precisión de la deserción del 66% para los árboles de decisiones y SVM.

En (Segura-Morales & Loza-Aguirre, 2018) buscan determinar cómo los factores socioeconómicos afectan los logros educativos de los estudiantes de secundaria. Se consideraron datos socioeconómicos y académicos correspondientes a más de diez años de registros obtenidos de la principal universidad de un país de la región andina. Utilizando algoritmos de clasificación y técnicas de aprendizaje automático para determinar qué factores influyen más en el rendimiento académico. Se encontró que las becas académicas, la edad, el condado y el grado de la escuela secundaria influyen en el rendimiento académico de los estudiantes.

Por otra parte, en (Solis et al., 2018) los autores analizaron el rendimiento de cuatro algoritmos de minería de datos con diferentes perspectivas para la definición del conjunto de datos, para la predicción de la deserción de estudiantes universitarios. La muestra está compuesta por todos aquellos estudiantes que se matricularon en un programa de grado en el Instituto Tecnológico de Costa Rica (ITCR) entre los años 2011 y 2016. Hubo 90.067 registros, correspondientes a las matrículas de 16.807 estudiantes, que inicialmente cumplieron este

criterio. Random Forest fue el algoritmo que obtuvo una predicción del 91% y una sensibilidad del 87%.

En (Pérez et al., 2019) hacen una comparación de indicadores de desempeño del modelo actual de deserción de la Universidad del Bío-Bío (UBB) en Chile, que se basa en la técnica de regresión logística y se compara con un nuevo modelo basado en árboles de decisión. La comparación muestra que la predicción de la deserción escolar del modelo propuesto obtiene una exactitud del 86%, una precisión del 97% con una tasa de error del 14%.

En (Beaulac & Rosenthal, 2019) analizan los dos primeros semestres de cursos completados por un estudiante para predecir si obtendrán un título de pregrado. Se analizaron un amplio conjunto de datos que contiene todos los cursos tomados por cada estudiante de pregrado en una de las principales universidades de Canadá durante 10 años. En este artículo, se construyeron dos clasificadores utilizando bosques aleatorios (Random forrest). Entre los estudiantes que completaron su programa en el conjunto de pruebas, el clasificador logra una precisión del 91,19%. De los 418 estudiantes que no completaron su programa, el clasificador logra una precisión del 52,95%. El resultado combinado es una precisión del 78,84% sobre el conjunto de pruebas completo.

En el siguiente estudio (Lee & Chung, 2019) el objetivo es mejorar el funcionamiento de un sistema de alerta temprana de deserción escolar, abordando el problema del desequilibrio de clase utilizando las técnicas de sobremuestreo de minorías sintéticas (SMOTE) y los métodos de conjunto en el aprendizaje automático. Evaluando los clasificadores capacitados con curvas tanto de características operativas del receptor (ROC) como de precision–recall (PR). Utilizando las muestras de grandes datos de los 165.715 estudiantes de secundaria del Sistema Nacional de Información Educativa (NEIS) de Corea del Sur. Se entrenaron a cuatro clasificadores: bosque

aleatorio (RF), árbol de decisión impulsado (BDT), bosque aleatorio con SMOTE (SMOTE + RF), y árbol de decisión impulsado con SMOTE (SMOTE + BDT). El árbol de decisión impulsado mostró el mejor desempeño.

Clasificadores y métricas utilizadas

Durante el desarrollo de las investigaciones sobre la deserción estudiantil se han utilizado distintos clasificadores y métricas, destacando en cuanto a los clasificadores el uso de árboles de decisión y en cuanto a las métricas que permiten valorar la calidad del modelo, la exactitud (Accuracy). La exhaustiva revisión de la literatura científica, en este ámbito de conocimiento, ha permitido identificar los clasificadores más frecuentemente usados en las investigaciones relacionadas, la tabla 4 contiene tal documentación.

Arboles de Decisiones - AD (Decision Tree), Maquinas de Soporte Vectorial – SVM (Support Vector Machine), Naive Bayes – NB, Regresión Logística – RL (Logistic Regression), Regresión Lineal – RLN (Linear Regression), Redes Neuronales Artificiales – ANN (Artificial Neural Networks), K vecinos más próximos - KNN (K-Nearest-Neighbor), CM (C-MEAN), Longitud mínima de descripción – MDL (Minimum Description Length).

Tabla 3.

Clasificadores usados para el análisis de la deserción estudiantil

Referencias	AD	SVM	NB	RL	RLN	RNA	KNN	CM	MDL
(Thomas & galambos, 2004)	X				X				
(Tsai et al., 2011; veitch, 2004)	X								
(Kalles & pierrakeas, 2006a; s. Kotsiantis et al., 2004)	X	X	X	X		X	X		
(S. B. Kotsiantis & pintelas, 2005)	X	X			X	X			
(Beaulac & Rosenthal, 2019; Hasbun et al., 2016; Kalles & Pierrakeas, 2006b; Lee & Chung, 2019; Márquez-Vera et al., 2013; T.	X								

Mishra et al., 2014; Moseley & Mead, 2008; Mustafa et al., 2012; Pereira et al., 2013; Segura-Morales & Loza-Aguirre, 2018)									
(Barker et al., 2004; Güner et al., 2014; Lykourentzou et al., 2009)		X				X			
(Askinadze & Conrad, 2017; Sangodiah et al., 2015)		X							
Referencias	**AD**	**SVM**	**NB**	**RL**	**RLN**	**RNA**	**KNN**	**CM**	**MDL**
(Cheewaprakobkit, 2013)	X					X			
(Alkhasawneh & Hobson, 2011; Cambruzzi et al., 2015; Jin et al., 2011; Lesinski et al., 2016; Sarker et al., 2014; Siri, 2015)						X			
(*Predicting Students Drop Out A Case Study*, 2009)	X		X	X					
(Bayer et al., 2012; Şara et al., 2015)	X	X	X						
(Barbosa Manhães et al., 2014; Manhães et al., 2014; Santana et al., 2015)	X	X	X			X			
(Delen, 2010; Solis et al., 2018; Zeng et al., 2017)	X	X		X		X			
(Zhang & Oussena, 2010)	X	X	X		X		X		X
(Salazar et al., 2004)	X							X	
(Delen, 2011; Hernandez Gonzalez et al., 2016)	X			X		X			
(Chai & Gibson, 2015; Pérez et al., 2019)	X			X					
(Miranda & Guzmán, 2017; Sharabiani et al., 2014)	X		X			X	X		
(Devasia et al., 2016; Krishna Kishore et al., 2014)	X		X			X			
(Aziz et al., 2015)			X						
(Dharmawan et al., 2018)	X	X					X		

Fuente: Elaboración propia

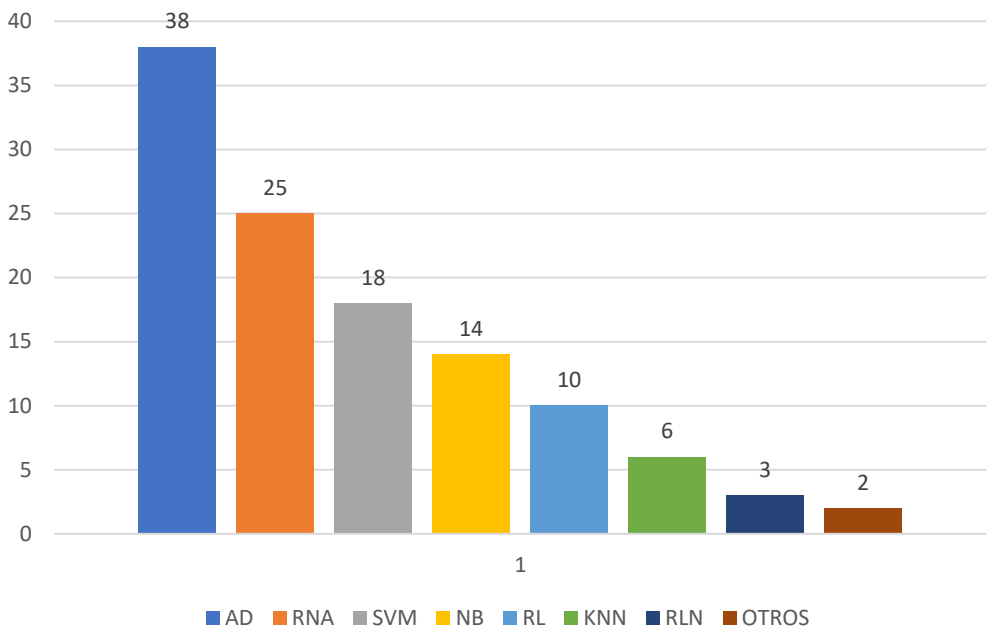


Figura 8. Frecuencia de uso de clasificadores en investigaciones consultadas

Fuente: Elaboración propia

En la tabla 4 se presentan las métricas usados en las investigaciones desarrolladas basadas en la literatura examinada. Exactitud – AC (Acurracy), Sensibilidad – SE (Sensibility), Especificidad – ES (Specificity), Media Geométrica – GM (Geometric Media), Cobertura – RC (Recall), Verdaderos Negativos – TN (True Negatives), Falsos Positivos – FP (False Positives), F-Measure – FM, Medida-F – MF, Elevación – EL (Elevation), Otros – OT.

Tabla 4.

Métricas de Evaluacion usados para el análisis de la deserción estudiantil

Referencias	AC	SE	ES	ROC	GM	RC	TN	FP	FM	MF	EL	OT
(Bayer et al., 2012; kalles & pierrakeas, 2006a; moseley & mead, 2008; sarker et al., 2014; thomas & galambos, 2004)	X	X	X									
(Aziz et al., 2015; beaulac & rosenthal, 2019; cambruzzi et al., 2015; cheewaprakobkit, 2013; delen, 2011; dharmawan et al., 2018; hasbun et al., 2016; jin et al., 2011; kalles & pierrakeas, 2006b; s. B. Kotsiantis & pintelas, 2005; mustafa et al., 2012; pereira et al., 2013; sharabiani et al., 2014; siri, 2015; tsai et al., 2011; veitch, 2004; zhang & oussena, 2010)	X											
(Alkhasawneh & hobson, 2011; bayer et al., 2012; delen, 2010; lykourentzou et al., 2009; manhães et al., 2014)	X	X										
(Márquez-vera et al., 2013)	X	X	X		X							
(T. Mishra et al., 2014)	X	X				X						
(Barbosa manhães et al., 2014; santana et al., 2015)	X	X	X				X	X				
(Krishna kishore et al., 2014)	X	X	X	X		X			X			
(Miranda & guzmán, 2017)	X	X	X	X		X		X	X			
(Askinadze & conrad, 2017; chai & gibson, 2015; sangodiah et al., 2015)	X					X						
(Şara et al., 2015; zeng et al., 2017)	X			X								
(Lesinski et al., 2016)	X		X			X						
(Hernandez gonzalez et al., 2016)	X	X								X	X	
(Pérez et al., 2019; segura-morales & loza-aguirre, 2018)	X	X	X				X					X
				X		X						

Fuente: Elaboración propia

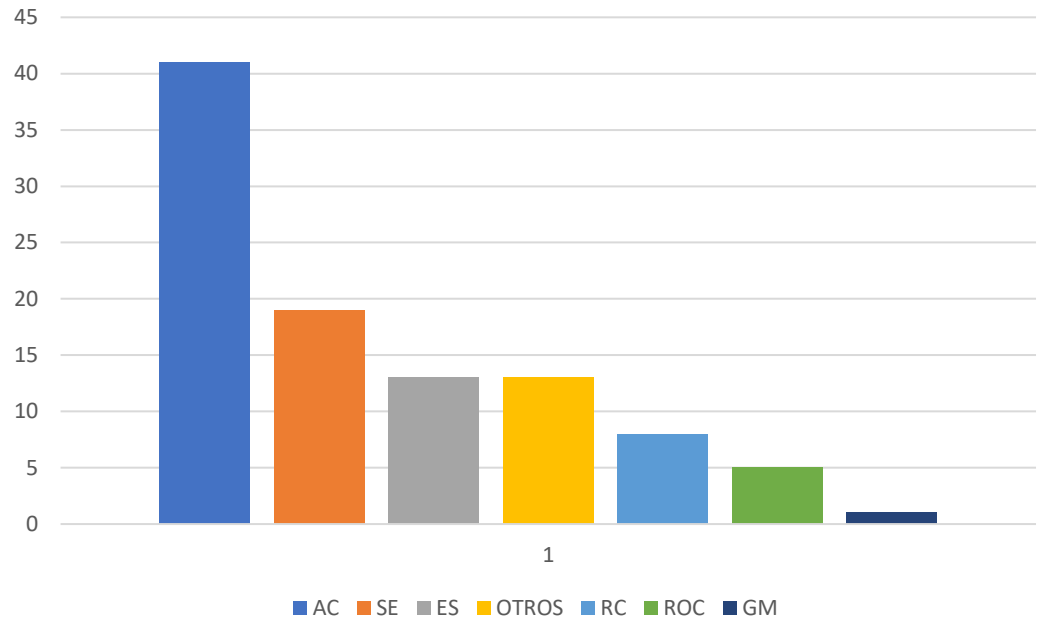


Figura 9. Frecuencia de uso métricas en investigaciones consultadas

Fuente: Elaboración propia

Identificación de problemas comunes

Basándose en el análisis literario realizado, se observa la importancia de determinar las causas que generan la deserción estudiantil. Los autores utilizan distintas fuentes de información para poder crear los conjuntos de datos, modelos y algoritmos, con el propósito de detectar los estudiantes en riesgo de desertar. En la búsqueda de esto, se han podido identificar varios desafíos y retos, tales como:

- El sobre entrenamiento (Overfitting) es un problema causado cuando un clasificador identifica una estructura que corresponde demasiado con el conjunto de entrenamiento y se generaliza mal a las nuevas observaciones (Alkhasawneh & Hobson, 2011; Barnes et al., 2009; Beaulac & Rosenthal, 2019; Lee & Chung, 2019; Lykourentzou et al., 2009).

- Datos faltantes del conjunto de datos (Moseley & Mead, 2008; Sangodiah et al., 2015; Thomas & Galambos, 2004).
- Conjunto de datos desbalanceados (Delen, 2010; Moseley & Mead, 2008; Sarker et al., 2014). Cuando hay un porcentaje alto de estudiantes que permanecen y pocos datos de estudiantes que han desertado.
- Dificultad para obtener la información debido a la confidencialidad de los mismos (S. Kotsiantis et al., 2004; Lykourentzou et al., 2009; Sangodiah et al., 2015; Zhang & Oussena, 2010).

Modelo para la predicción de la deserción de estudiantes de pregrado, basado en técnicas de minería de datos

En este proyecto de investigación se propone un modelo para predecir la deserción de estudiantes de pregrado de la Universidad de la Costa - CUC, basado en la implementación de técnicas de minería de datos. El objetivo final es detectar lo más precozmente posible a los estudiantes que presentan alto riesgo de desertar, con el fin de proporcionar algún tipo de ayuda para tratar de evitar y/o reducir el fracaso estudiantil en los primeros semestres.

Metodología utilizada

El método utilizado en este trabajo para predecir la deserción estudiantil es el de descubrimiento de conocimientos en bases de datos – KDD, descrito en detalle en la sección La figura 1 los describe de forma mucho mas entendible.

Las principales etapas del método, aplicadas a esta investigación, son descritas detalldamente a continuación:

- Recopilación de datos: Se obtiene toda la información disponible sobre los estudiantes. Para ello, el conjunto de factores que pueden afectar el rendimiento de los estudiantes debe ser identificado y tomado de las diferentes fuentes de datos disponibles. Toda la información debe integrarse en un conjunto de datos. En la sección 4.2 se detalla todo el proceso realizado para la recopilación de los datos.
- Preprocesamiento de datos: El conjunto de datos en esta etapa está preparado para aplicar las técnicas de minería de datos. Aplicando métodos de preprocesamiento, como limpieza de datos, transformación de variables, selección de atributos y rebalanceo de los datos. Estos últimos para resolver los problemas de alta dimensionalidad y desequilibrio de los

datos que suelen presentarse en los conjuntos de datos. En la sección 4.3 se detallan los pasos realizados para el preprocesamiento de los datos.

- Minería de datos e interpretación: En esta fase, se utilizan algoritmos de minería de datos para predecir el fracaso de los estudiantes como un problema de clasificación. Se ejecutan, evalúan y comparan diferentes algoritmos para determinar cuál de ellos obtiene los mejores resultados. Se analizan los modelos obtenidos para detectar la deserción estudiantil. En la sección 4.4 se detalla el proceso de minería de datos.

Recopilación de datos

En este proyecto de investigación se ha usado la información de estudiantes de pregrado de la Universidad de la Costa CUC, de los periodos comprendidos entre 2013-1 y 2018-2. Esta se recolecta a partir de la información registrada en los formatos de inscripción de los estudiantes. Ellos registran allí sus antecedentes demográficos, culturales, sociales, familiares y educativos, estatus socioeconómico y algunos aspectos relacionados con su perfil psicológico. Esta información es compilada por el departamento de Bienestar Universitario, mediante el Programa para el Acompañamiento y Seguimiento para la Permanencia Estudiantil – PASPE.

El conjunto de datos recopilado está constituido por 23 características o columnas de datos, cuya descripción se presenta en la Tabla 5 (indicando el tipo de dato por cada característica) y por 10.939 registros o instancias de datos únicas, correspondientes a información descriptiva de cada estudiante. Las características del presente estudio fueron comparadas con las utilizadas en otros estudios afines. Para ello, se requirió el análisis de 20 artículos de investigación. La Tabla 6 contiene la identificación de las característcas usadas en otros estudios afines, para analizar los procesos de deserción estudiantil.

Tabla 5.

Descripción de las características del conjunto de datos propuesto

No	Característica	Tipo de dato	Comentarios
1	Programa académico	Categórico	Se refiere al programa académico que está cursando. Por ejemplo: Ingeniería de Sistemas, Ingeniería Electrónica, Administración de empresas, etc.
2	Procedencia	Categórico	Es el lugar donde habitualmente reside. Es decir, la ciudad de donde procede, por ejemplo: Barranquilla, Campo de la cruz, etc.
3	Labora	Lógico (boleano)	Indica si está o no trabajando. Sus posibles valores son: si o no.
4	Tiene hijos	Lógico (boleano)	Indica si tiene hijos o no. Sus posibles valores son: si o no.
5	Estrato	Numérico	Se refiere al estrato socioeconómico. Su rango de valores es de 1 a 6.
6	Tipo de colegio	Categórico	Indica si la institución educativa escolar es pública o privada.
7	Discapacidad	Lógico (boleano)	Indica si tienen alguna discapacidad. Sus posibles valores son: si o no
8	Sexo	Categórico	Indica al sexo que pertenece. Sus posibles valores son: masculino o femenino
9	Edad	Categórico	Se refiere al rango de edad en el cual está comprendido. Los posibles valores son: rango1, rango2, rango3, rango4, rango5, rango6, rango7, rango8 o rango9.
10	Estado civil	Categórico	Indica si está casado, soltero, separado o en unión libre
11	Ocupación madre	Categórico	Indica si labora, ama de casa, pensionada, fallecida, estudiante o madre cabeza de hogar
12	Ocupación padre	Categórico	Indica si labora, pensionado, fallecido o esta desempleado
13	Posee computador	Lógico (boleano)	Indica si tiene computador. Sus posibles valores son: si o no
14	Acceso a internet	Lógico (boleano)	Indica si tiene Internet en su casa. Sus posibles valores son: si o no
15	Posee celular inteligente	Lógico (boleano)	Indica si tiene celular inteligente o Smartphone Sus posibles valores son: si o no
16	Tiene plan de datos	Lógico (boleano)	Indica si tiene plan de datos en su celular. Sus posibles valores son: si o no
17	Número de hermanos	Categórico	Indica la cantidad de hermanos que tiene. Sus posibles valores son: 1, 2, 3, 4, 5, 6 o más de 7
No	**Característica**	**Tipo de dato**	**Comentarios**
18	Posición hermanos	Categórico	Indica la posición que ocupa dentro de sus hermanos. Sus posibles valores son: 1, 2, 3, 4, 5, 6 o más de 6
19	Número integrantes familia	Categórico	Indica el número de familiares que hacen parte del núcleo familiar. Sus posibles valores son: 1, 2, 3, 4, 5, 6, 7, más de 7

20	Minoría	Categórico	Indica si pertenece alguna minoría étnica. Sus posibles valores son: etnia indígena, afrodescendiente, víctima del conflicto, no pertenece
21	Talento o capacidad excepcional	Lógico (boleano)	Indica si posee algún talento o capacidad excepcional. Sus posibles valores son: si o no
22	Afiliado a EPS	Lógico (boleano)	Indica si está afiliado a una EPS. Sus posibles valores son: si o no
23	Usa lentes recetados	Lógico (boleano)	Indica si usa lentes recetados. Sus posibles valores son: si o no

Fuente: Elaboración propia

Tabla 6.

Análisis de las características empleadas en estudios de deserción estudiantil afines

Referencias	**Características**
(Ahuja & kankane, 2017)	nombre de la escuela, sexo del estudiante, edad del estudiante, domicilio del estudiante, tamaño de la familia, estado de cohabitación de los padres, calificación educativa de la madre, calificación educativa del padre, trabajo de la madre, trabajo del padre, tiempo de viaje de la casa a la escuela, tiempo de estudio semanal, número total de fracasos de clases anteriores, clases extra (matrícula) para cualquier materia del curso, actividades extracurriculares, acceso a Internet en casa, cualquier relación romántica, tiempo libre después de la universidad, tiempo con amigos, consumo de alcohol durante la semana, consumo de alcohol durante los fines de semana, estado de salud actual, número de literas universitarias, primer grado del semestre, segundo grado del semestre, calificación final del semestre (todo numérico, de 0 a 20), initgrado (resultado de la escuela secundaria superior), drop (caída tomada por el estudiante o no), Dstatus (atributo resultante). Dstatus es el resultado de la finalización de la carrera de un estudiante
(Hoffait & schyns, 2017)	género, nacionalidad y campo de estudio, agrupados en Ciencias Humanas, Ciencias y Ciencias de la Salud, fecha de nacimiento, escolaridad previa, la concesión o no de una beca
(Asif et al., 2017)	notas de preadmisión de los estudiantes Y las notas de todos los cursos que se imparten en los cuatro años de la carrera
(Peralta et al., 2017)	Promedio ponderado acumulado, Código de carrera que sigue alumno, Quintil del alumno, Año de matrícula, Categoría de solicitud de motivos de renuncia, Módulo de prueba de ciencia rendida, Puntaje asignado al promedio de notas, Puntaje de prueba matemática actual, Puntaje prueba lenguaje anterior, Vía de ingreso, Centro de estudios previo de alumno, Número de personas que componen el grupo familiar, Promedio ponderado en último año, Puntaje ponderado, Situación de aprobación o desaprobación, Plan de estudio, Estado Civil, Beca académica
(Kumar baradwaj & pal, 2011)	Calificaciones del semestre anterior, calificación de la prueba de clase, desempeño en el seminario, asignación, competencia general, asistencia, trabajo de laboratorio, calificaciones al final del semestre
(Bayer et al., 2012)	género, año de nacimiento, año de admisión, año de admisión, exención del examen de ingreso, puntuación en el examen de capacidad de estudio, número de semestres terminados, cursos reconocidos, créditos reconocidos, créditos a obtener, créditos ganados, cursos no terminados, segundas restas hechas, días excusados, calificaciones promedio, calificaciones promedio ponderadas, la proporción del número de créditos ganados al número de créditos a ganar, la diferencia entre los créditos ganados y los créditos a el número de estudios paralelos en la facultad, el número de estudios paralelos en la universidad, el número de todos los estudios en la facultad, el número de todos los estudios en la universidad

Referencias	**Características**
(Márquez-vera et al., 2016)	Promedio de notas en la escuela secundaria, puntuación media en EXANI I, Aula/grupo inscrito, tamaño de la clase, edad, asistencia durante las sesiones de la mañana/noche, nivel de ingresos de la familia, tener una beca, tener un trabajo, vivir con los padres, nivel de educación de la madre y nivel de educación del padre. Tener una discapacidad física, altura, peso, cintura, medida de flexibilidad, ejercicios abdominales en un minuto, flexiones en un minuto, tiempo en una carrera de 50 m, tiempo en una carrera de 1000 m, consumo regular y regular de

	alcohol y hábitos de fumar. Asistencia, nivel de aburrimiento durante las clases, mala conducta y sanción administrativa, número de amigos, número de horas diarias de estudio, horas de estudio en grupo, lugar de estudio habitual, hábitos de estudio, forma de resolver dudas, nivel de motivación, religión, influencia externa en la elección de la titulación, tipo de personalidad, recursos para el estudio, número de hermanos/hermanas, posición como el hijo mayor/medio/joven, el fomento del estudio por parte de los padres, el número de años que viven en la ciudad, el método de transporte utilizado para ir a la escuela, la distancia a la escuela, el interés por las asignaturas, el nivel de dificultad de las asignaturas, la toma de notas en clase, la demanda excesiva de deberes, los métodos de enseñanza, la calidad de la infraestructura escolar, el hecho de contar con un tutor personal y el nivel de preocupación del profesor por el bienestar de cada estudiante Puntuación en Matemáticas, puntuación en Física, puntuación en Ciencias Sociales, puntuación en Humanidades, puntuación en Escritura y Lectura, puntuación en Inglés y puntuación en Ciencias de la Computación Que abandonan o continúan en el próximo semestre
(Pradeep et al., 2015)	Número de estudiantes en clase, asistencia durante las sesiones de la mañana y la noche, discapacidad física, enfermedad grave, número de amigos, número de horas diarias de estudio, hábitos de estudio, estudio en grupo, fomento del estudio por parte de los padres, estado civil, métodos de estudio, recursos para el estudio, religión, tipo de personalidad, nivel de ingresos de la familia, tener una beca, vivir con los padres, nivel de educación de la madre, nivel de educación del padre, vivir en una gran ciudad, nivel de motivación, tomar notas en clase, métodos de enseñanza, número de años viviendo en la ciudad, método de transporte utilizado para ir a la escuela, distancia a la escuela, consumo regular de alcohol, hábitos de fumar, nivel de asistencia durante las clases, nivel de aburrimiento durante las clases, interés en las asignaturas, nivel de dificultad de las asignaturas, demanda excesiva de deberes, calidad de la infraestructura escolar, tener un tutor personal, número de hermanos/hermanas, posición como el niño más grande/medio/joven.
(Fernandes et al., 2019)	Coordinación regional de la educación, Región administrativa de la escuela, Escuela, Turno, Clase con personas con necesidades especiales, Entorno de uso del aula, Código de estudiante, Género, Edad (media), Beneficio estudiantil, Ciudad estudiantil, Barrio estudiantil, Estudiante con necesidades especiales, Materias escolares, Grado (media), Ausencia (media), Resultado final del estudiante
(Shahiri et al., 2015)	Evaluaciones internas, Factores psicométricos, Evaluación externa, Demografía estudiantil, Antecedentes de la escuela secundaria, Becas, Interacción con redes sociales, Evaluación interna, Actividades extracurriculares, Demografía estudiantil, Antecedentes de la escuela secundaria
(Quadri & kalyankar, 2010)	Género, Asistencia, Semestres anteriores, Parentesco, Ingresos de los padres, Becas, Primer hijo, Trabajo, Deserción escolar
(Osmanbegovi, 2012)	Género, Distancia, GPA, Becas, Materiales, Grado importancia, Familia, Bachillerato, Examen de ingreso, Tiempo, Internet, Ganancias
(Tair, 2015)	Identificación del estudiante, Nombre del estudiante, Sexo, Fecha de nacimiento, Lugar de nacimiento, Especialidad, Año de inscripción, Año de graduación, Ciudad, Ubicación, Dirección, Teléfono, Matriculación GPA, Desgarro de matriculación, GPA de la universidad, Grado (Excelente, Bueno o Promedio)
(Mayilvaganan & kalpanadevi, 2015)	Identificación del estudiante, Nombre del estudiante, Sexo, Fecha de nacimiento, Lugar de nacimiento, Especialidad, Año de inscripción, Año de graduación, Ciudad, Ubicación, Dirección, Teléfono, Matriculación GPA, GPA de la universidad, Grado (Excelente, Bueno o Promedio)
(Christian & ayub, 2014)	Género, Facultad, Departamento, Admisión, Puntuación del examen, Ciudad, Especialidad, GPA, Condición, Departamento de Estudiantes, Tipo de fase de admisión
(Heredia et al., 2015)	Semestre en el cual está matriculado el estudiante, Edad actual, Ciudad de procedencia, Estrato, Jornada, Sexo, Valor de la matrícula, Ocupación, Materias cursadas, Materias perdidas, Promedio, Estado Civil, Nivel de estudios del Padre, Nivel de estudios de la Madre, Ingresos y Desertado (Atributo de clase)
(Devasia et al., 2016)	Género, Categoría de los estudiantes, Medio de enseñanza, Hábito alimentario de los estudiantes, Otro hábito de los estudiantes , Lugar de vida, Dónde se aloja , Número de miembros de la familia, Situación familiar de los estudiantes, Situación económica anual de la familia, Grado de los estudiantes en el 10º / SSLC, Grado Estudiantil en 12º/ PUC, Tipo de Colegio Estudiantil, Calificación del Padre, Calificación de la Madre, Ocupación del Padre, Ocupación de la Madre, Estudiante Interesado en la Educación Superior, ¿Utiliza Móvil? Si la respuesta es Sí. Desde cuántos Meses/Años, Internet, red social, Cuántos Hermanos y su cualificación, Hábito de lectura, ¿Cuántas horas al día dedica a los estudios?
(Dharmawan et al., 2018)	Género, Distancia del hogar, Estado de vivienda, Estado civil, Número de miembros de la familia, Tiempo de espera para estudiar, Internet en la casa, Intensidad de uso del teléfono móvil, Interés de salud en estudios posteriores, Interés en las especialidades, Motivación del estudio, Expectativa de idoneidad para las especialidades, Estatus de empleo, Educación de la madre, Estatus de Empleo de la Madre, Educación del Padre, Estatus de Empleo del Padre, Relación con los estudiantes, Relación con la Familia, Relación con los Profesores, Logro, Deferencia, Orden, Exhibición, Autonomía, Afiliación, Dominio, Abandono, Crianza, Cambio, Resistencia, Heterosexualidad, Agresión, Consistente

Referencias	**Características**
(B. Perez et al., 2018)	Información de admisión, incluyendo información demográfica mínima (sexo, fecha de nacimiento, estado civil). Fechas de graduación, incluyendo la fecha de graduación y el programa académico. Expedientes académicos incluyendo los cursos tomados y las calificaciones de cada uno de ellos, el programa y el promedio académicos acumulativo. Ayudas financieras, incluyendo todas las ayudas financieras en los términos requeridos

(Gulati, 2015)	programa, idioma, fecha nacimiento, edad, sexo, empleado, años exp, horas trabajadas x semana, horario de trabajo, salario, religión, hermanos, vive con padres, casado, hijos, nacionalidad, lugar residencia, área (rural, urbana), estado, distancia de casa, modo transporte, tiempo de viaje

Fuente: Elaboración propia

Se analizó la frecuencia de uso de las características en los artículos de investigación, distinguiendo las que tenían la frecuencia de uso mayor o igual a tres (3). En la Tabla 7 se detalla la frecuencia de uso de las características usadas en los conjuntos de datos de los artículos de investigación consultados.

Tabla 7.

Frecuencia de uso de las características utilizadas para predecir la deserción estudiantil, a partir de los artículos de investigación consultados

Características	**Frecuencia de uso en artículos de investigación citados**
Genero	14
Promedio acumulado	9
Estado civil	7
Nivel estudios madre	7
Nivel estudios padre	7
Ciudad/ubicación	6
Distancia a la institución educativa	6
Fecha nacimiento	5
Edad	5
Trabaja	5
Número de hermanos	5
Colegio	4
Beca	4
Vive con sus padres	4
Nivel económico familiar	4
Tamaño familia	4
Tiempo de estudio semanal	4
Año matricula	3
Características	**Frecuencia de uso en artículos de investigación citados**
Semestres finalizados	3
Posee alguna discapacidad	3
Trabaja la madre	3

Trabaja el padre	3
Nivel de motivación	3
Tiene internet en casa	3
Método de transporte para ir a Estudiar	3
Consumo de alcohol	3
Consumo de alcohol fines de semana	3
Materias vistas	3

Fuente: Elaboración propia

Una de las limitantes de este estudio es que no todas las características indicadas en la tabla 7 fue posible obtenerlas, debido a que no fueron en su momento solicitadas en las entrevistas de los estudiantes, de 28 se obtuvieron 21 características de la Tabla 8. Sin embargo, fue posible obtener ocho (8) características o atributos adicionales tomados del sistema académico de la Universidad (puntaje ICFES, ingreso familiar, vivienda propia, número hermanos educación superior, validó bachillerato, reintegro, tipo de inscripción y cantidad de becas aplicadas), para un total de 32 características. La tabla 8 presenta una detallada descripción de la estructura final del conjunto de datos.

Tabla 8.

Características del conjunto de datos definitivo propuesto

No	Características	Valores
1	Programa académico	administración ambiental, administración de empresas, administración de servicios de salud, arquitectura, banca y finanzas, comunicación social y medios digitales, contaduría pública, derecho, ingeniería agroindustrial, ingeniería ambiental, ingeniería civil, ingeniería de sistemas, ingeniería eléctrica, ingeniería electrónica, ingeniería industrial, licenciatura en educación básica primaria, mercadeo y publicidad, negocios internacionales, psicología, administración de empresas virtual
2	Procedencia	barranquilla o municipio del atlántico, fuera del atlántico, fuera del país, no registra
No	**Características**	**Valores**
3	Labora	Si, no, no registra
4	Tiene hijos	Si, no, no registra
5	Estrato	1, 2, 3, 4, 5, 6
6	Tipo colegio	Público, privado, no registra

7	Discapacidad	Si, no, no registra
8	Sexo	Masculino, femenino
9	Edad	15-16, 17-18, 19-20, 21-22, 23-24, 25-26, 27-28, 29-30, más de 30
10	Estado civil	soltero, unión libre, casado, separado, no registra
11	Ocupación madre	labora, ama de casa, pensionada, fallecida, estudiante, madre cabeza de hogar, no registra
12	Ocupación padre	labora, pensionado, fallecido, desempleado, no registra
13	Posee computador	Si, no, no registra
14	Acceso a internet	Si, no, no registra
15	Posee celular inteligente	Si, no, no registra
16	Tiene plan de datos	Si, no, no registra
17	Número de hermanos	0,1,2,3,4,5,6,7, más de 7
18	Posición hermanos	0,1,2,3,4,5,6, más de 6, no registra
19	Número integrantes familia	1,2,3,4,5,6,7, más de 7, no registra
20	Minoría	pueblo indígena, comunidad afrodescendiente, víctima del conflicto, no registra, no pertenece
21	Talento o capacidad excepcional	Si, no, no registra
22	Afiliado a EPS	Si, no, no registra
23	Usa lentes recetados	Si, no, no registra
24	Puntajes ICFES	Numérico
25	Ingreso familiar	Numérico
26	Vivienda propia	Si, no
27	Número hermanos educación superior	0,1,2,3,4,5,6,7,10
28	Validó bachillerato	Si, no
29	Reintegro	Si, no
30	Tipo de inscripción	transferencia interna, normal, transferencia externa, reintegro, reserva cupo, exo transf interna, exo normal
31	Cantidad de becas aplicadas	Numérico
32	Desertó	Si, no

Fuente: Elaboración propia

Para determinar la cantidad de becas aplicadas, se identificaron en el sistema académico todas las becas a las que el estudiante aplicó y se contaron, es importante resaltar que no se sumaron los montos adjudicados por concepto de tales becas. En cuanto a los posibles valores que toma la características tipo de inscripción, el valor exo (exo transf interna y exo normal),

hace referencia a reserva de cupo sin pago de inscripción o que aún no ha cancelado su matrícula.

Para el etiquetado de los datos, se utilizó como criterio de clase la característica "desertó". Para ello, fue necesario identificar por cada estudiante, quienes habían desertado o no de la Universidad. Esto implicó un cruce de información del listado de estudiantes contenidos en el conjunto de datos recopilado, con el sistema académico SICUC1, el cual arrojó el estado de los estudiantes en el sistema (Activo, Inactivo, Egresado, Graduado, Anulado, Cancelado o Excluido). Los estudiantes que tienen el estado Activo, Egresado y Graduado, se etiquetaron en la característica desertó como "no", por otra parte, los que estaban como Inactivos, Anulado, Cancelado y Excluido, se etiquetaron en la característica desertó como "si", debido a que no se encontraban matriculados al momento de realizar la consulta.

Preprocesamiento de datos

En esta etapa, según (Márquez-Vera et al., 2013; Pereira et al., 2013), se deben realizar algunas tareas de preprocesamiento como la limpieza, integración, discretización y transformación de las características. Es importante destacar que una tarea muy importante en este proyecto de investigación ha sido el preprocesamiento de datos, debido a que la calidad y fiabilidad de la información disponible, afecta directamente a los resultados obtenidos. Se aplicaron algunas tareas específicas de preprocesamiento, para preparar todos los datos descritos anteriormente con el fin de que la tarea de clasificación pudiera llevarse a cabo correctamente. En primer lugar, todos los datos disponibles se integraron en un único conjunto de datos. Durante este proceso se eliminaron los estudiantes que no tenían el 100% de la información completa,

[1] Es la plataforma en línea de gestión académica que soporta los procesos de la Corporación Universidad de la Costa CUC

arrojando un total de 1606 registros de estudiantes para ser analizados en el conjunto de datos, con una reducción en los datos del 85%.

Adicionalmente, se realizaron modificaciones a los valores de algunas características. Por ejemplo, las palabras que contenían la letra "ñ" fueron reemplazadas por "n". Se modificó el atributo de la edad de cada estudiante reemplazando por los valores que se presentan en la tabla 9.

Tabla 9.

Valores para reemplazo en la característica edad

Rango de edades	**Valor para reemplazo**
15-16	Rango1
17-18	Rango2
19-20	Rango3
21-22	Rango4
23-24	Rango5
25-26	Rango6
27-28	Rango7
29-30	Rango8
Mas de 30	Rango9

Fuente: Elaboración propia

Además, se sustituyeron las tildes del conjunto de datos por vocales sin tildes y los espacios en blanco fueron reemplazados por “_” guion de piso, esto incluyó también los nombres de las características del conjunto de datos. Una vez concluida la transformación de los datos, se procedió a guardar el conjunto de datos en el formato.arff, compatible con la herramienta WEKA. Dicha herramienta es un programa de código abierto, empleada para realizar diferentes procesos propios de la minería de datos. WEKA ha sido desarrollado en Java por la Universidad de Waikato en Nueva Zelanda (Birjali et al., 2018).

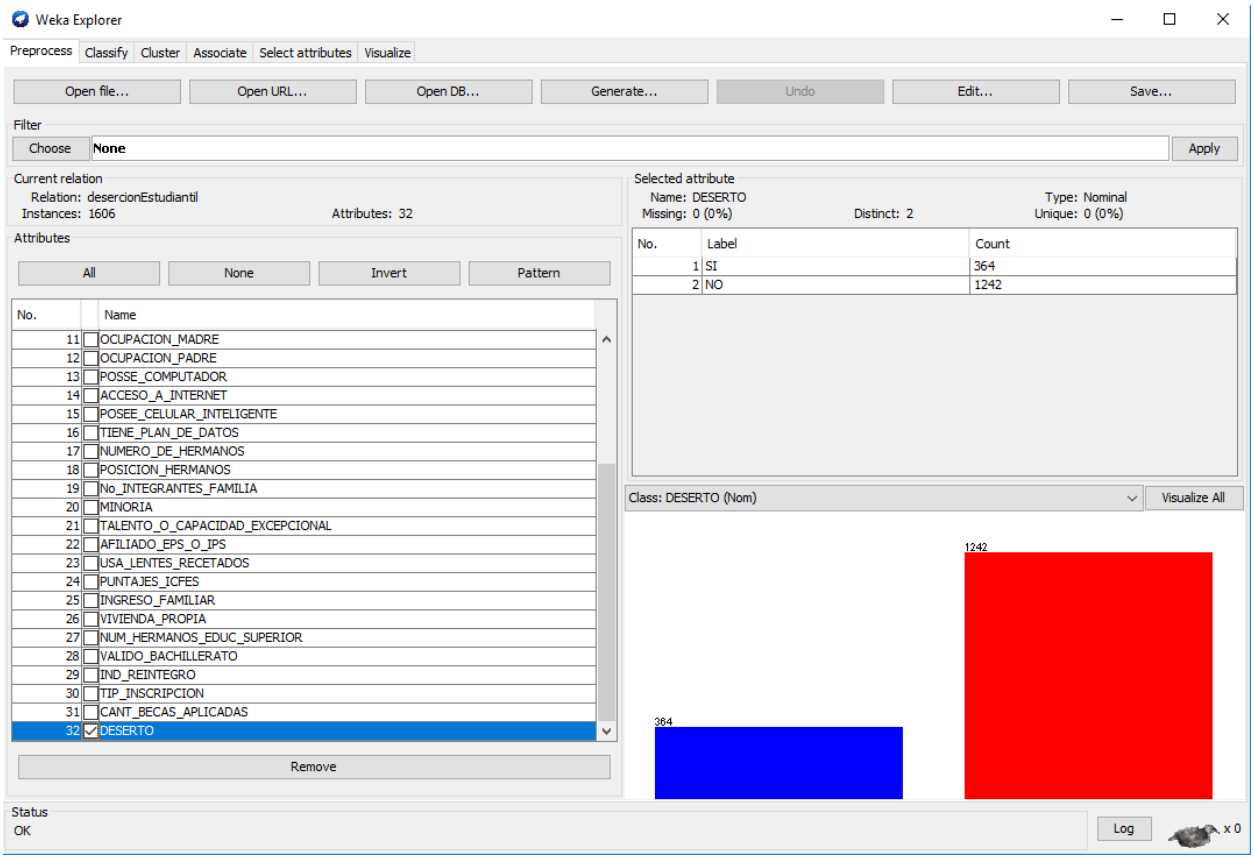


Figura 10. Conjunto de datos en formato arff cargado en WEKA

Fuente: Elaboración propia

Como se puede apreciar en la figura 10, el conjunto de datos está desbalanceado, es decir, hay más estudiantes que no han desertado (1242) que los que han desertado (364). Según (Márquez-Vera et al., 2013), el problema de la clasificación desbalanceada de datos ocurre cuando el número de instancias en una clase es mucho menor que el número de instancias en otra clase u otras clases. Una forma de solucionar esto durante la etapa de preprocesamiento de los datos, es realizando un muestreo o balanceando las clases (Márquez-Vera et al., 2013). En este proyecto de investigación se utilizó la técnica de sobre muestreo de minorías sintéticas, para el balanceo de clases (Synthetic Minority Oversampling Technique – SMOTE), con ello se soluciona el desbalanceo de los datos, esta técnica ha sido utilizada en varios artículos de

investigación (Han et al., 2012; Lee & Chung, 2019; Márquez-Vera et al., 2013). Se seleccionó SMOTE, con *nearestNeighbors* 15 y *percentage* 240. Se balancearon los registros de la característica desertó, si (1237) y no (1242).

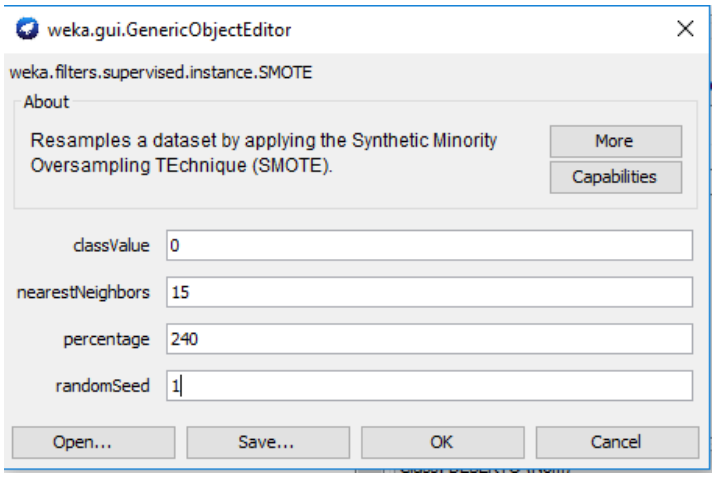


Figura 11. Parámetros SMOTE seleccionados para balanceo del atributo desertó

Fuente: Elaboración propia

Al aplicar los ajustes en WEKA, tal como se evidencian en la figura 12, se logra balancear los registros e instancias del conjunto de datos, respecto a la característica o la característica “deserto”.

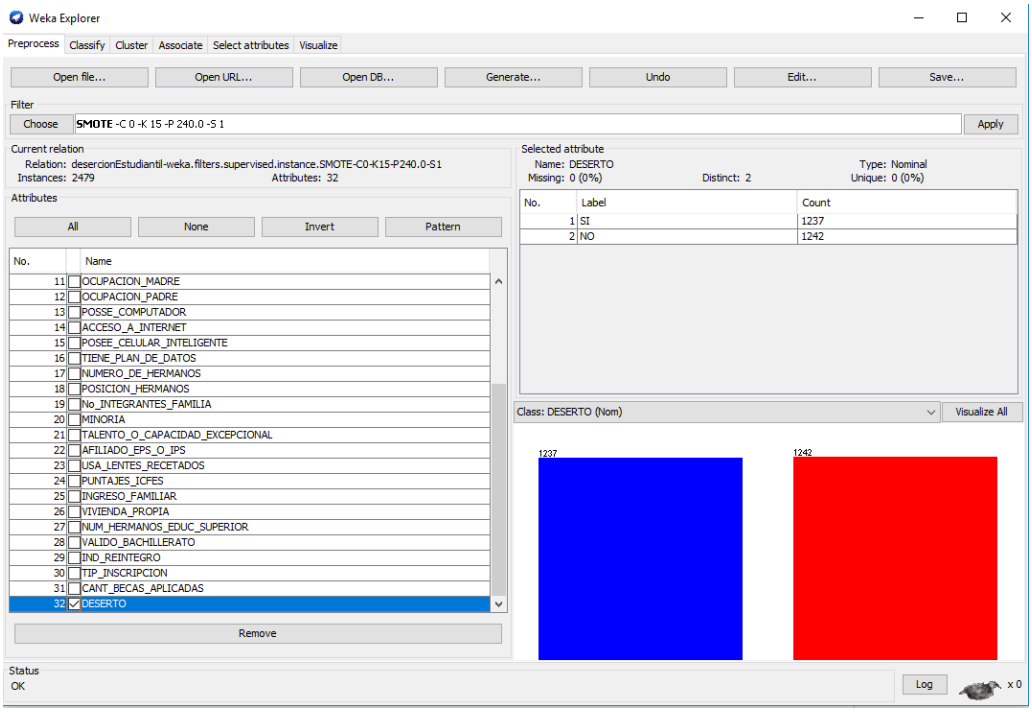


Figura 12. Datos balanceados con SMOTE

Fuente: Elaboración propia

Minería de datos

En esta etapa se detallan los diferentes escenarios de experimentación efectuados, usando un variado repertorio de técnicas de Machine Learning, con el propósito de obtener el modelo que mejor prediga cuales estudiantes están en riesgo de desertar. Se realizaron tres escenarios de experimentación con el fin de obtener la más alta exactitud (acurracy), en el primer escenario se aplicaron las técnicas de clasificación utilizando el conjunto de datos desbalanceado, en el segundo escenario se aplicaron las técnicas de clasificación con el conjunto de datos balanceado (producto de la implementación de la técnica de balanceo SMOTE). Es importante resaltar que en los dos primeros escenarios se utilizaron las 31 características del conjunto de datos. En el tercer escenario se identificaron características que agregaban ruido al proceso y por tanto fueron eliminadas del *dataset*, luego se balanceo el *dataset* y se efectuó un proceso de prueba (test)

mucho más exhaustivo. A continuación, se muestran los resultados obtenidos, producto de la ejecución del primer escenario de experimentación, utilizando el *dataset* sin balancear, aplicando la técnica de prueba (*test*) validación cruzada (*cross-validation*) con 10 pliegues.

Tabla 10.

Resultados de la clasificación utilizando el conjunto de datos sin balancear

Técnica de clasificación	TP Rate	FP Rate	Precision	Accuracy	F-Measure	ROC Area
BayesNet	76.5%	56.8%	73.6%	76.5%	74.4%	69.2%
NaiveBayes	73.8%	53.1%	72.6%	73.8%	73.2%	67%
NaiveBayesSimple	73.5%	52.8%	72.5%	73.5%	73%	66.8%
NaiveBayesUpdateable	73.8%	53.1%	72.6%	73.8%	73.2%	67%
SMO	77.8%	72.5%	74.1%	77.8%	70.2%	52%
ADTree	78.1%	64.7%	74.3%	78.1%	73.4%	71.5%
RandomForest	75.7%	70.4%	68.9%	75.7%	70%	60.1%
J48	77.5%	62.9%	73.6%	77.5%	73.6%	61.1%

Fuente: Elaboración propia

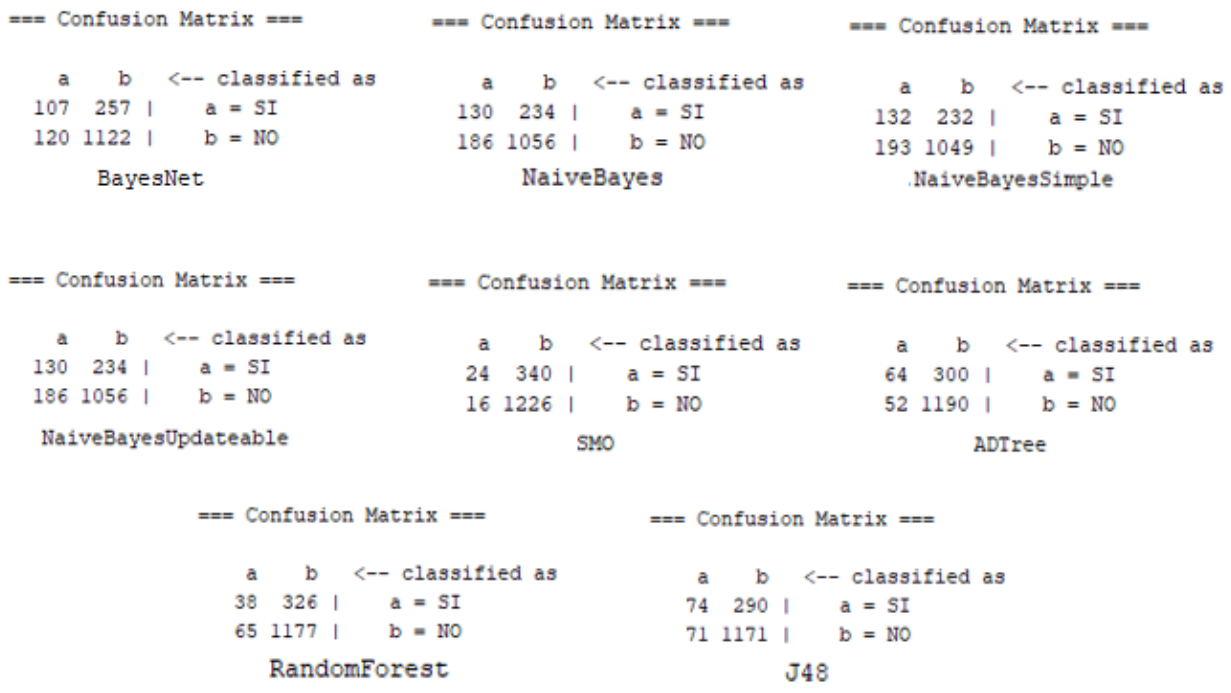


Figura 13. Matriz de confusión datos sin balancear

Fuente: Elaboración propia

En la figura 13 se observan las matrices de confusión de los calificadores utilizados al realizar el proceso con los datos sin balancear (BayesNet, NaiveBayes, NaieBayesSimple, NaiveBayesUpdatable, SMO, ADTree, RandomForest y J48. Así mismo, en la figura 14 se toma una simulación de Weka con el clasificador BayesNet, donde se observan todas las métricas arrojadas y la matriz de confusión.

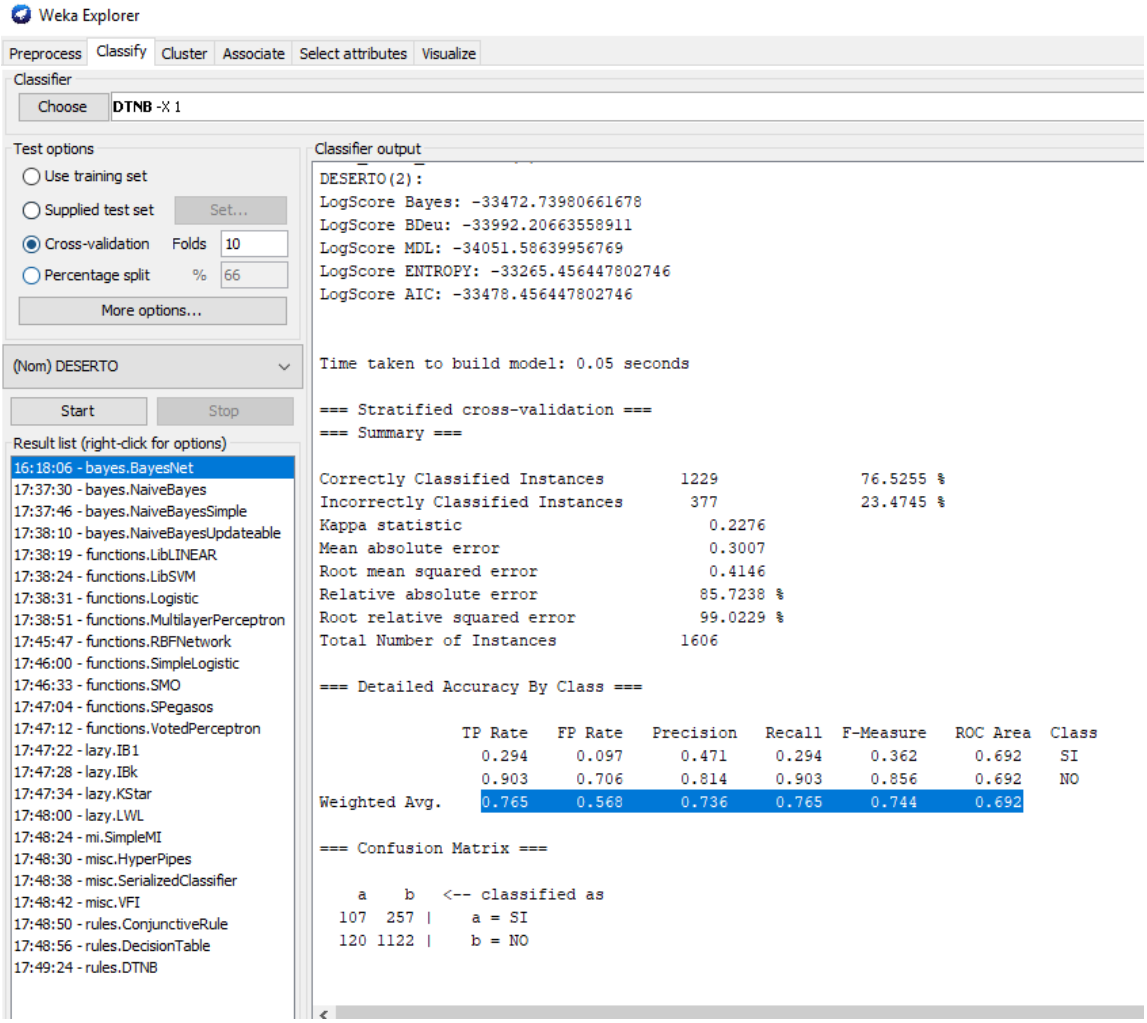


Figura 14. Clasificadores utilizados en el conjunto de datos desbalanceado

Fuente: Elaboración propia

En cuanto al segundo escenario de experimentación, en la tabla 10 se muestran los resultados obtenidos a partir de la aplicación de diferentes técnicas de Machine Learning al

conjunto de datos sin balancear, tales técnicas son: clasificadores bayesianos (BayesNet, NaiveBayes, NaiveBayesSimple y NaiveBayesUpdateable), máquinas de soporte vectorial – SVM (cuya denominación en la herramienta WEKA es SMO) y diferentes técnicas basadas en árboles de decisión (ADTree, RandomForest y J48). La métrica de exactitud obtenida oscila entre el 73% y el 78%, evidenciando que la técnica de clasificación con mejores prestaciones fue ADTree con una exactitud (accuracy) del 78.1%. Es importante resaltar que la métrica ROC area, para este clasificador equivale al 71.5%, valor cercano al 50.0% lo cual denota aleatoriedad.

En la siguiente etapa se realizó el mismo proceso utilizando el conjunto de datos balanceados, aplicando para ello la técnica SMOTE y la técnica de prueba (test) validación cruzada (cross-validation) con 10 pliegues, los resultados obtenidos se presentan en la tabla 11.

Tabla 11.

Resultados de la clasificación utilizando el conjunto de datos balanceados con cross-validation a 10 pliegues

Técnica de clasificación	TP Rate	FP Rate	Precision	Accuracy	F-Measure	ROC Area
BayesNet	84%	16.1%	85.3%	84%	83.8%	86.9%
NaiveBayes	74.9%	25.1%	75%	74.9%	74.8%	84.3%
NaiveBayesSimple	74.9%	25%	75.1%	74.9%	74.9%	84.4%
NaiveBayesUpdateable	74.9%	25.1%	75%	74.9%	74.8%	84.3%
SMO	80.4%	19.6%	80.4%	80.4%	80.4%	80.4%
ADTree	77.2%	22.9%	77.5%	77.2%	77.1%	84.2%
RandomForest	84.3%	15.7%	85.7%	84.1%	84.2%	88.9%
J48	79.6%	20.4%	79.8%	79.6%	79.5%	81.9%

Fuente: Elaboración propia

En la tabla 11 se observa el reporte comparativo, producto de la aplicación de las diferentes técnicas de clasificación antes mencionadas, utilizando el conjunto de datos

balanceado. Se observa un aumentó en la precisión, la cual osciló entre 75.0% y 85.7%. La técnica de clasificación con mejores prestaciones fue RandomForest (la cual construye múltiples árboles de decisión) con una exactitud del 84.1%. Es importante resaltar que la métrica ROC área, para este clasificador equivale al 88.9%, evidenciando un incremento porcentual importante en relación a las pruebas efectuadas con el conjunto de datos desbalanceado. Las mejoras en cuanto a las métricas de calidad, al usar el clasificador RandomForest con un conjunto de datos balanceado, según (Ahuja & Kankane, 2017) se deben a que: 1) los árboles se construyen al azar, superponiendo conjuntos de datos y características; 2) la selección aleatoria junto con los conjuntos superpuestos elimina el problema del ajuste excesivo u overfitting y 3) la decisión combinada o media sobre un gran número de árboles resulta en la eliminación de errores individuales y limitaciones de los árboles de decisión.

En la figura 15 se observan las matrices de confusión de los calificadores utilizados al realizar el proceso con los datos balanceados y Cross-validation a 10 pliegues (BayesNet, NaiveBayes, NaieBayesSimple, NaiveBayesUpdatable, SMO, ADTree, RandomForest y J48. Así mismo, en la figura 16 se toma una simulación de Weka con el clasificador RamfomForest, donde se observan todas las métricas arrojadas y la matriz de confusión.

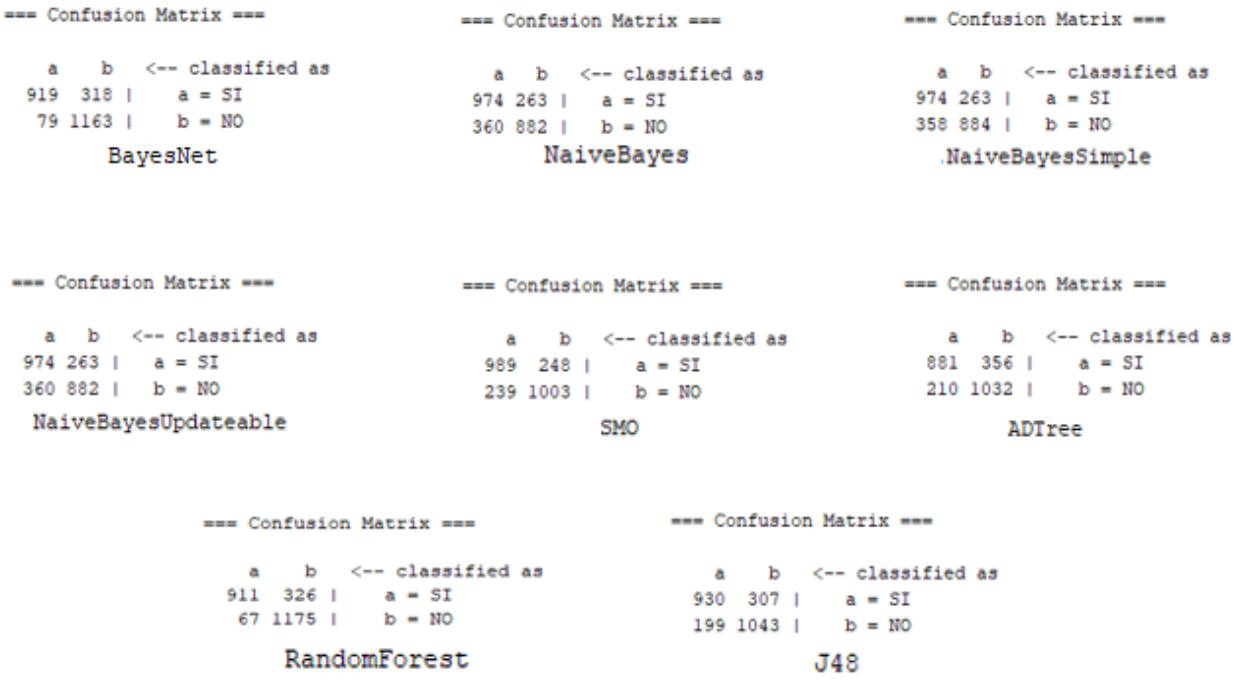


Figura 15. Matriz de confusión obtenida al aplicar la técnica RandomForest con datos balanceados y Cross-validation a 10 pliegues

Fuente: Elaboración propia

En un tercer escenario de experimentación, se buscó mejorar la precisión eliminando algunas características que no tienen tanta incidencia en la clasificación y efectuando un test mucho más exhaustivo. Posterior a ello, se hizo el balanceo de los datos con SMOTE configurado con los siguientes parámetros: nearestNeighbors 15 y percentage 240, el conjunto de datos resultante quedó constituido por un total de 2479 registros, de los cuales 1237 corresponden al criterio de clase “si desertaron” y 1242 al criterio de clase “no desertaron”. Las características eliminadas fueron (becas aplicadas y estado civil), debido a que, al momento de realizar varias simulaciones, tratando de identificar características sin peso, estas afectaban un poco de forma positiva los resultados finales, en especial el *accuracy*. En cuanto al proceso de pruebas (test) se realizaron las simulaciones con los clasificadores antes mencionados, en todos los casos aplicando cross-validation con 25 pliegues, los resultados de este escenario de simulación se presentan en la tabla 12.

Tabla 12.

Resultados de la clasificación utilizando el conjunto de datos balanceado y eliminación de algunas características aplicando cross-validation a 25 pliegues

Técnica de clasificación	TP Rate	FP Rate	Precision	Accuracy	F-Measure	ROC Area
BayesNet	81.9%	18.1%	82.4%	81.9%	81.9%	85.8%
NaiveBayes	73.8%	26.2%	74%	73.8%	73.7%	82.7%
NaiveBayesSimple	73.9%	26.1%	74.2%	73.9%	73.8%	82.8%
NaiveBayesUpdateable	73.8%	26.2%	74%	73.8%	73.7%	82.7%
SMO	78.3%	21.7%	78.3%	78.3%	78.3%	78.3%
ADTree	75.5%	24.5%	75.8%	75.5%	75.4%	81.5%
RandomForest	84.8%	15.2%	86.4%	84.8%	84.7%	88.2%
J48	80.3%	19.7%	80.8%	80.3%	80.2%	81.1%

Fuente: Elaboración propia

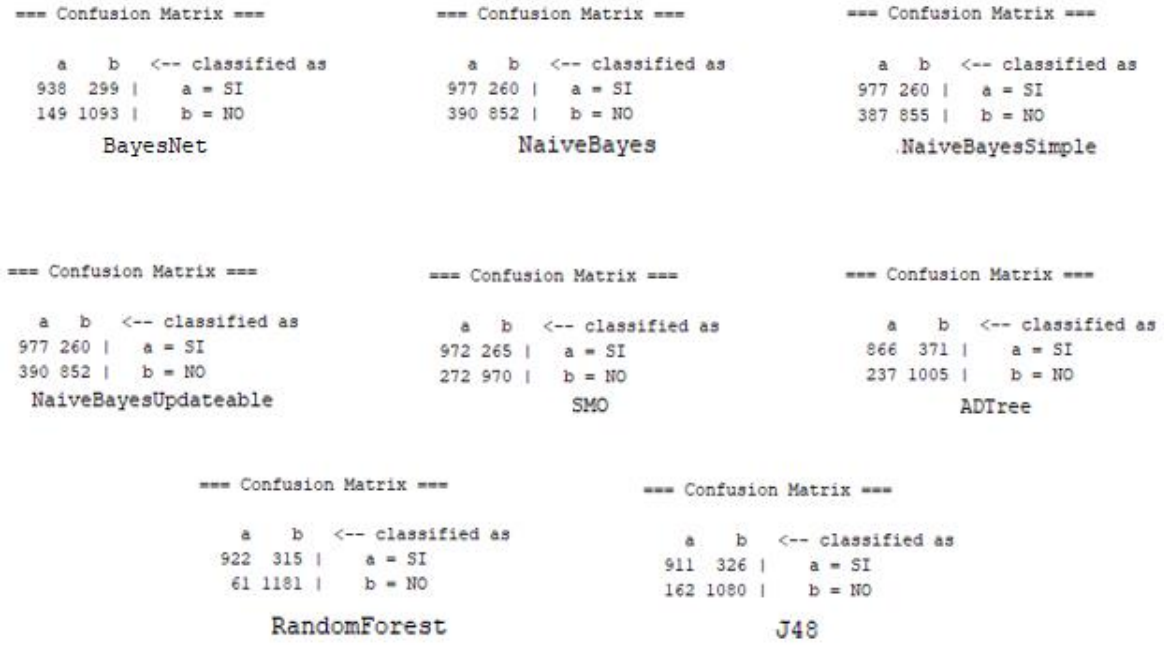


Figura 16. Matriz de confusión obtenida mediante datos balanceados y eliminación de algunas características, aplicando cross-validation a 25 pliegues

Fuente: Elaboración propia

En la figura 16 se observan las matrices de confusión de los calificadores utilizados al realizar el proceso con los datos balanceados, eliminación de algunas características y Cross-validation a 25 pliegues (BayesNet, NaiveBayes, NaieBayesSimple, NaiveBayesUpdatable,

SMO, ADTree, RandomForest y J48. Así mismo, en la figura 17 se toma una simulación de Weka con el clasificador RamfomForest, donde se observan todas las métricas arrojadas y la matriz de confusión.

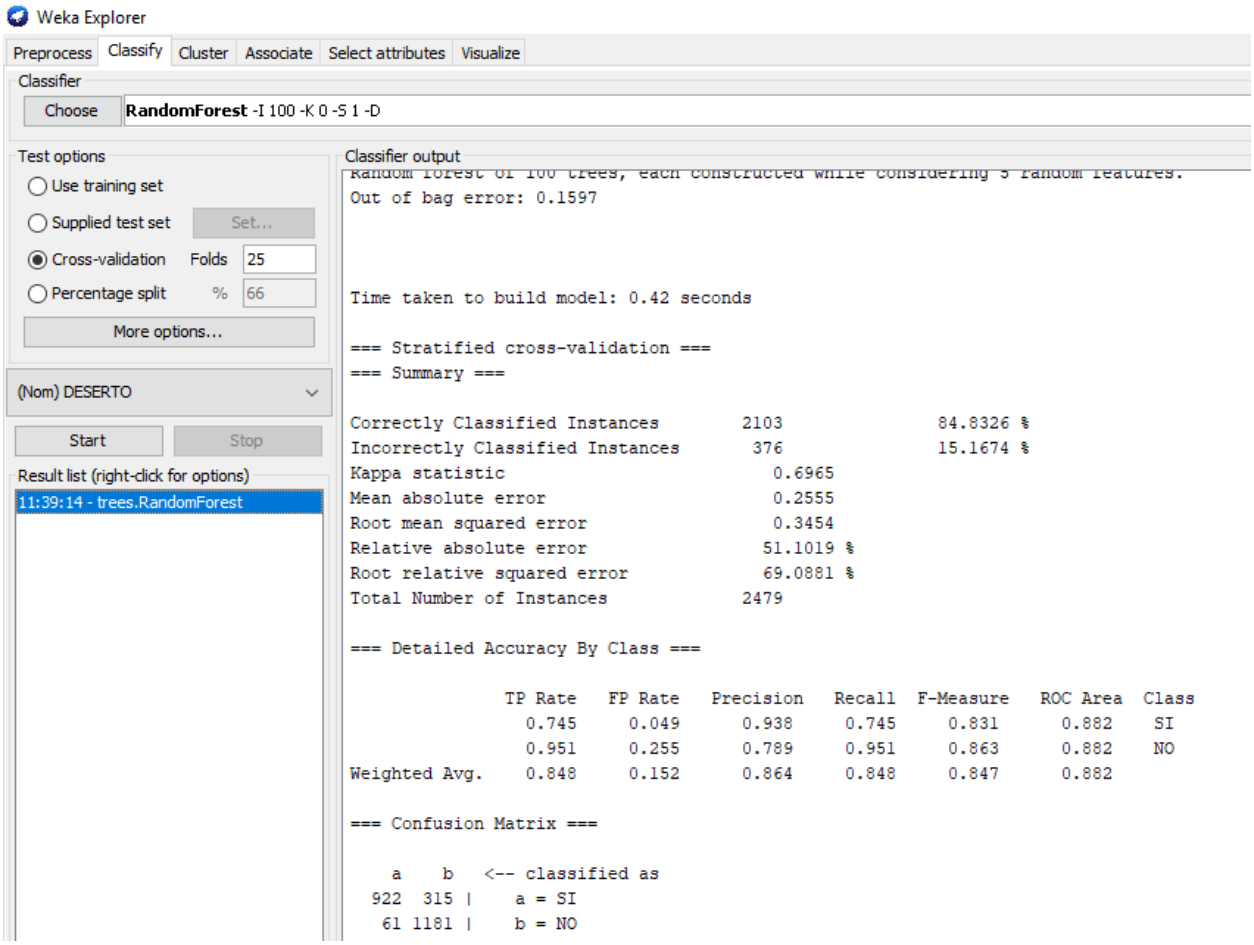


Figura 17. Simulación en WEKA con RandomForrest usando datos balanceados y eliminación de algunas características, con cross-validation a 25 pliegues

Fuente: Elaboración propia

Producto de la ejecución del escenario de experimentación No 3 y tomando como referente el escenario anterior, se logró aumentar la exactitud usando el clasificador RandomForrest, de 84.1% a 84.8%, con un valor en la métrica área ROC, bastante cercano al anterior escenario, es decir, de 88.9% a 88.2%. El número de pliegues equivalente a 25, para el

proceso de pruebas (test) efectuadas usando cross-validation, se determinó producto de un análisis de cada una de las iteraciones ejecutadas, lo cual permitió identificar con que cantidad de pliegues se obtenia el mas alto rendimiento de las métricas usadas.

Aplicación prototipo

El desarrollo del software prototipo fue basado en el conjunto de datos creado, implementando la mejor técnica de minería de datos obtenida de este proyecto, la cual fue RamdonForrest. A continuación, se observa el diagrama de flujo del aplicativo:

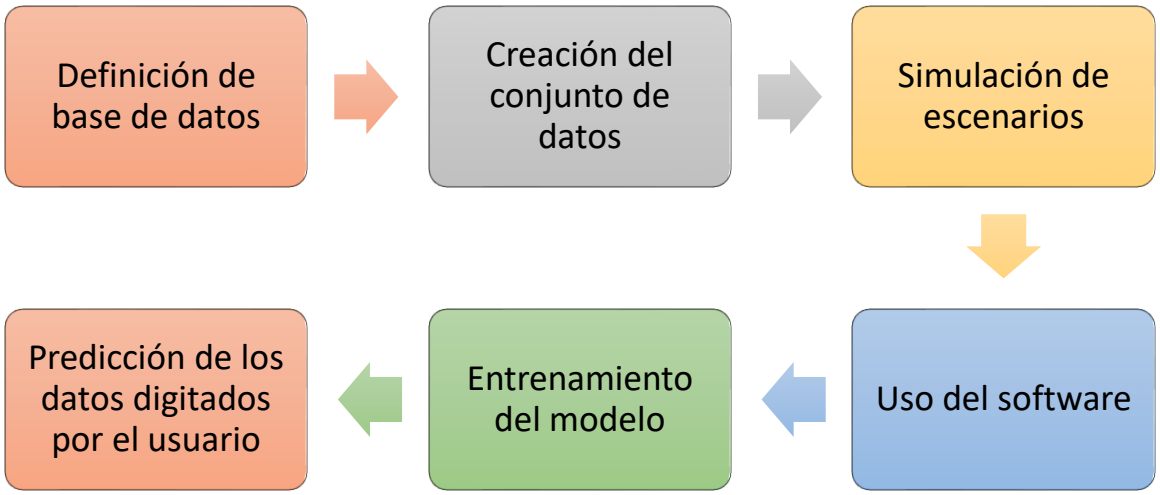


Figura 18. Diagrama de flujo para el desarrollo del aplicativo prototipo

Fuente: Elaboración propia

Proceso de desarrollo

Para el desarrollo del prototipo se utilizó el software Apache Netbeans IDE, basado en Java. Se importó la librería weka.jar al proyecto, en la Figura 19 se puede ver la importación de la biblioteca en la herramienta utilizada. En la Figura 20 se observa la importación del modelo de RamdonForrest en Weka.

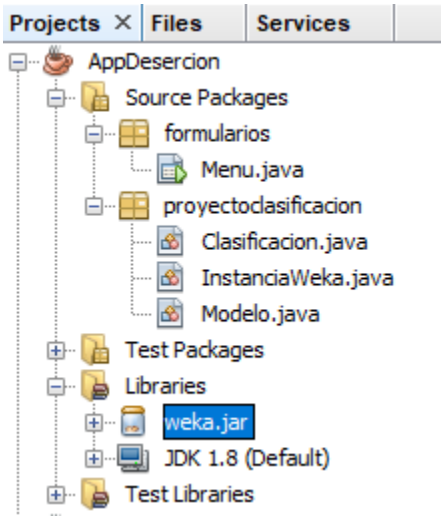


Figura 19. Uso de la librería Weka (weka.jar)

Fuente: Elaboración propia

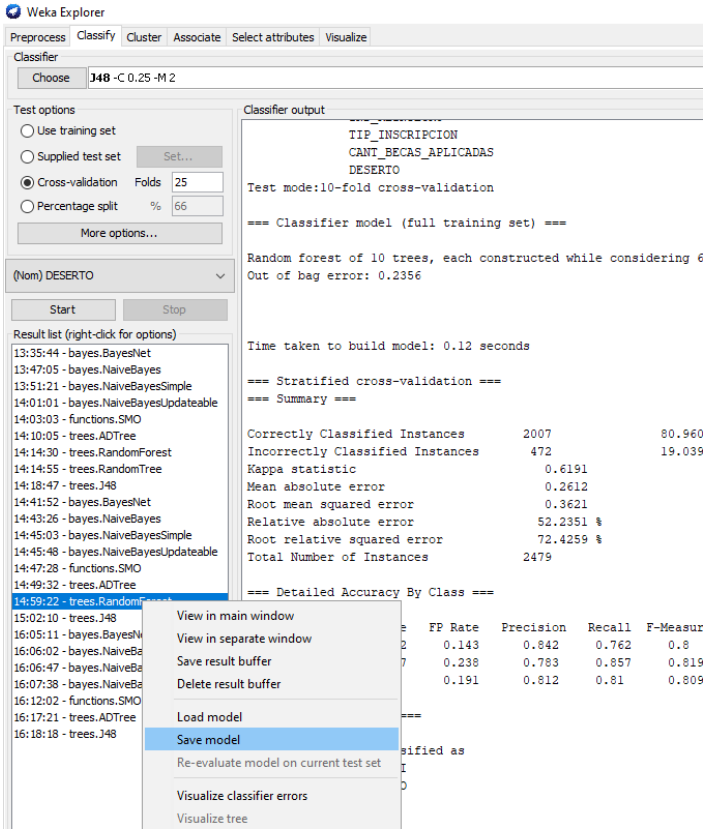


Figura 20. Exportar modelo en Weka

Fuente: Elaboración propia

Una vez importada la librería, se procedió a crear el formulario, métodos, procedimientos y clases. Primero se diseñó la distribución de la pantalla inicial, con las distintas opciones y distribución de la información requerida (Ver Fig. 21).

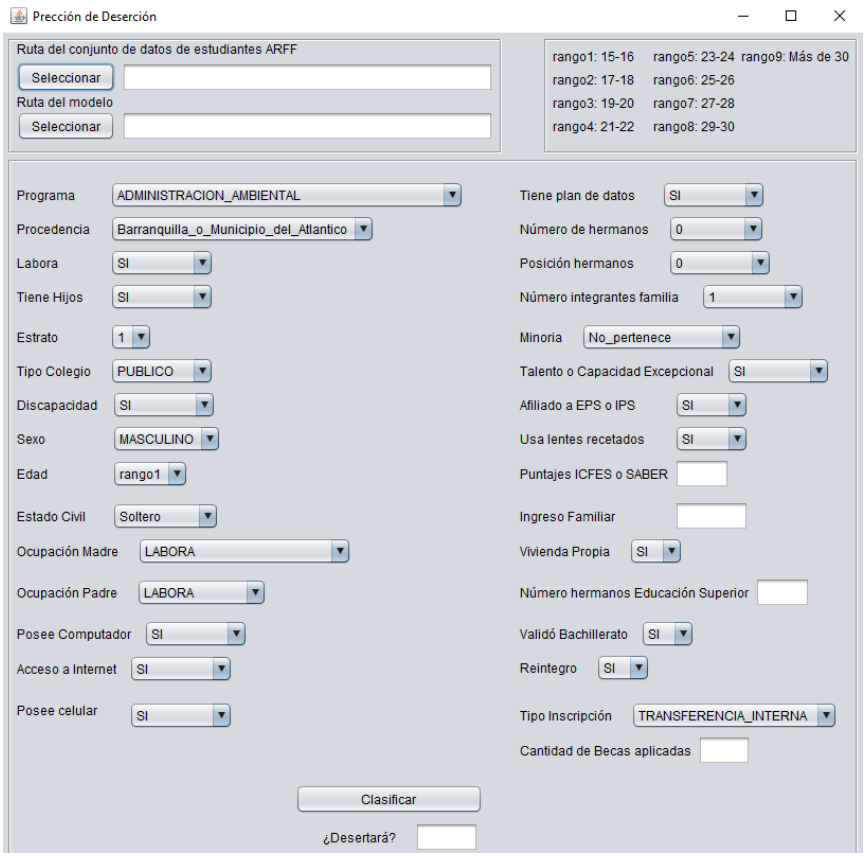


Figura 21. Formulario para predecir el riesgo de deserción de un estudiante

Fuente: Elaboración propia

En la figura 22 se muestra la clase Clasificación, donde se crea la instancia y se inicializa la clase de Weka RandomForest con los atributos obtenidos de la figura 21, y se obtiene la clasificación en un String al ejecutar el clasificador.

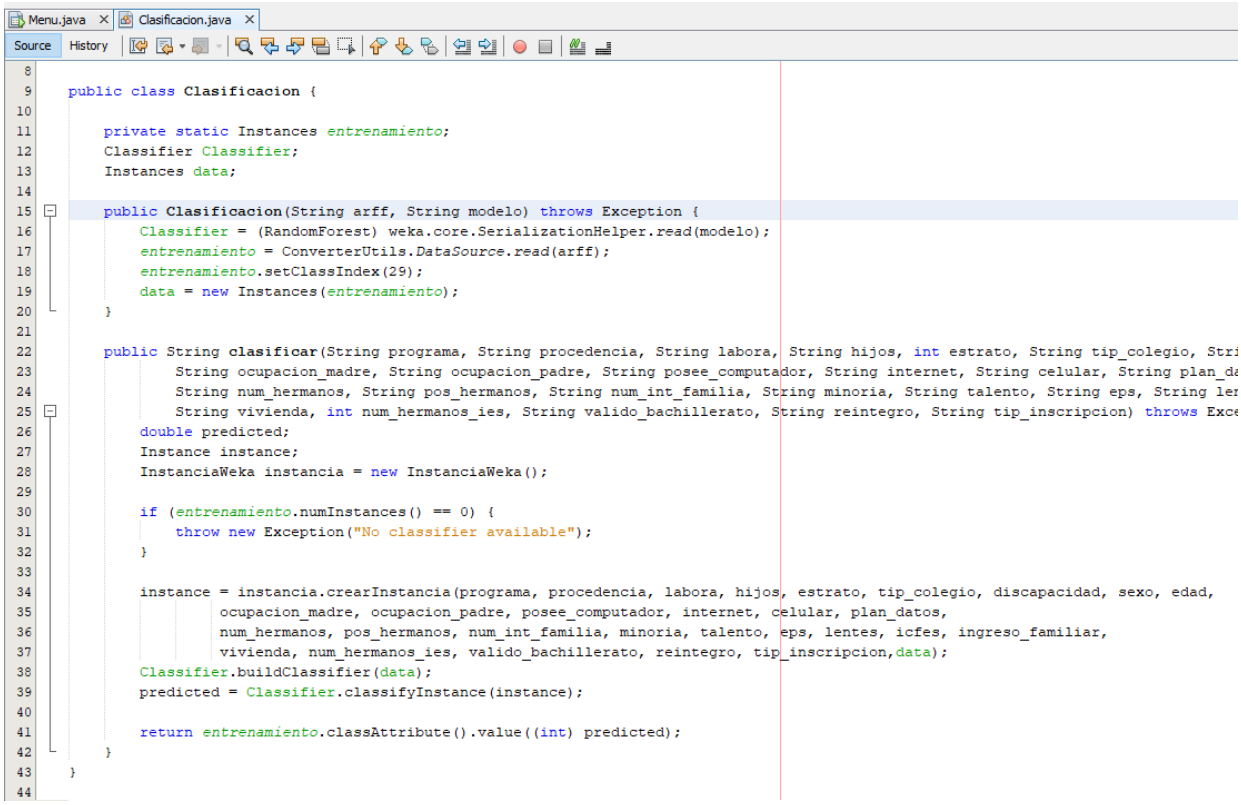


Figura 22. Código fuente clasificación Weka

Fuente: Elaboración propia

Arquitectura de la aplicación

La aplicación es stand-alone, todo se ejecuta en una maquina, cargando el .jar y las librerías requeridas para su ejecución. Se debe tener el JRE (Java SE Runtime Environment) en el PC para ejecutar el jar. Se descarga de la siguiente url:

- https://www.oracle.com/java/technologies/javase-jre8-downloads.html

Para el desarrollo de la aplicación se utilizaron las librerías dom4j-1.6.1, poi-3.9-20121203, poi-excelant-3.9-20121203, poi-ooxml-3.9-20121203, poi-ooxml-schemas-3.9-20121203, poi-scratchpad-3.9-20121203 y xmlbeans-2.3.0 (figura 24), todas para poder trabajar con archivos Excel (xlsx). En la figura 23 se observa la estructura de la aplicación Netbeans, con la estructura por defecto creado por el Netbeans.

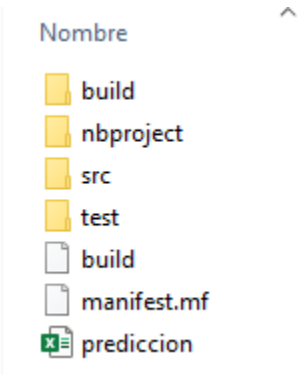


Figura 23. Estructura aplicación

Fuente: Elaboración propia

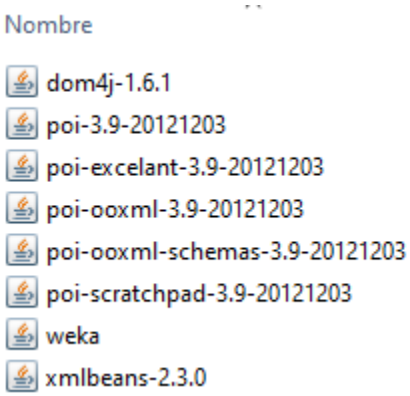


Figura 24. Librerías usadas en el aplicativo

Fuente: Elaboración propia

En la figura 25 se observa el diagrama de uso de la aplicación prototipo. Donde se puede realizar la predicción de un solo estudiante, seleccionando cada uno de los atributos requeridos y la predicción masiva, facilitando la predicción con gran cantidad de datos de forma automática utilizando un archivo Excel.

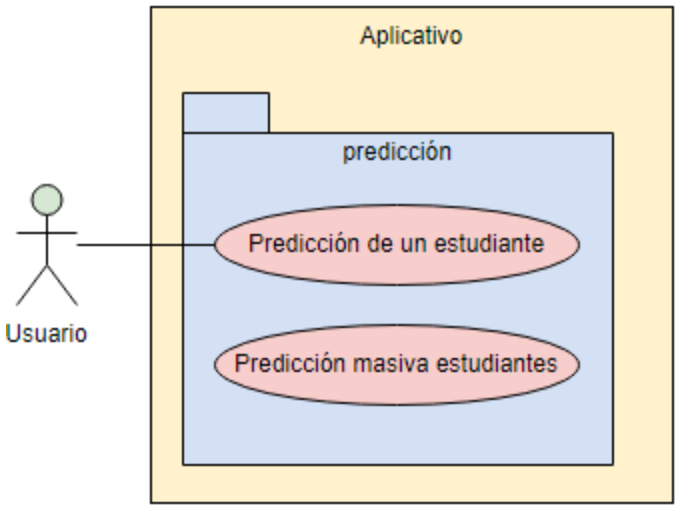


Figura 25. Diagrama de uso

Fuente: Elaboración propia

Descripción funcional de la aplicación

El usuario encontrará toda la información requerida para determinar si el estudiante estará en riesgo de desertar. Los datos fueron validados para evitar datos faltantes, haciéndolos mandatorios, para mejorar la predicción. La aplicación debe cargar el conjunto de datos Arff y el modelo entrenado, este es cargado seleccionado los archivos arrf y model. En la figura 26 se selecciona el archivo arff y el model.

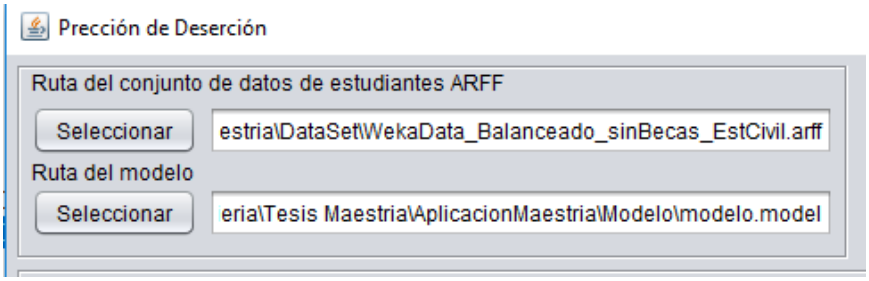
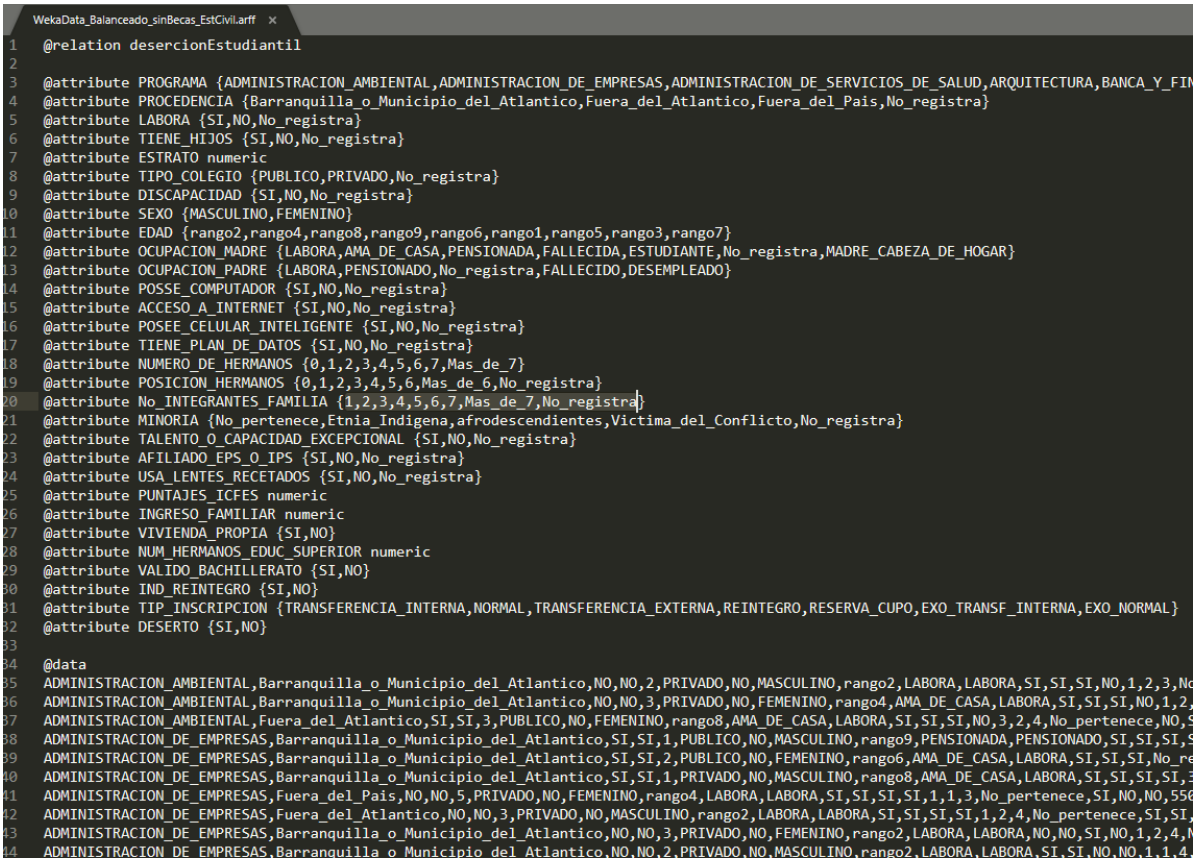


Figura 26. Ruta del conjunto de datos y el modelo entrenado

Fuente: Elaboración propia

A continuación, puede encontrar el archivo ARFF que se utilizó para generar el modelo y entrenar la herramienta, como se muestra en la Figura 27.



```
WekaData_Balanceado_sinBecas_EstCivil.arff
@relation desercionEstudiantil

@attribute PROGRAMA {ADMINISTRACION_AMBIENTAL,ADMINISTRACION_DE_EMPRESAS,ADMINISTRACION_DE_SERVICIOS_DE_SALUD,ARQUITECTURA,BANCA_Y_FIN
@attribute PROCEDENCIA {Barranquilla_o_Municipio_del_Atlantico,Fuera_del_Atlantico,Fuera_del_Pais,No_registra}
@attribute LABORA {SI,NO,No_registra}
@attribute TIENE_HIJOS {SI,NO,No_registra}
@attribute ESTRATO numeric
@attribute TIPO_COLEGIO {PUBLICO,PRIVADO,No_registra}
@attribute DISCAPACIDAD {SI,NO,No_registra}
@attribute SEXO {MASCULINO,FEMENINO}
@attribute EDAD {rango2,rango4,rango8,rango9,rango6,rango1,rango5,rango3,rango7}
@attribute OCUPACION_MADRE {LABORA,AMA_DE_CASA,PENSIONADA,FALLECIDA,ESTUDIANTE,No_registra,MADRE_CABEZA_DE_HOGAR}
@attribute OCUPACION_PADRE {LABORA,PENSIONADO,No_registra,FALLECIDO,DESEMPLEADO}
@attribute POSSE_COMPUTADOR {SI,NO,No_registra}
@attribute ACCESO_A_INTERNET {SI,NO,No_registra}
@attribute POSEE_CELULAR_INTELIGENTE {SI,NO,No_registra}
@attribute TIENE_PLAN_DE_DATOS {SI,NO,No_registra}
@attribute NUMERO_DE_HERMANOS {0,1,2,3,4,5,6,7,Mas_de_7}
@attribute POSICION_HERMANOS {0,1,2,3,4,5,6,Mas_de_6,No_registra}
@attribute No_INTEGRANTES_FAMILIA {1,2,3,4,5,6,7,Mas_de_7,No_registra}
@attribute MINORIA {No_pertenece,Etnia_Indigena,afrodescendientes,Victima_del_Conflicto,No_registra}
@attribute TALENTO_O_CAPACIDAD_EXCEPCIONAL {SI,NO,No_registra}
@attribute AFILIADO_EPS_O_IPS {SI,NO,No_registra}
@attribute USA_LENTES_RECETADOS {SI,NO,No_registra}
@attribute PUNTAJES_ICFES numeric
@attribute INGRESO_FAMILIAR numeric
@attribute VIVIENDA_PROPIA {SI,NO}
@attribute NUM_HERMANOS_EDUC_SUPERIOR numeric
@attribute VALIDO_BACHILLERATO {SI,NO}
@attribute IND_REINTEGRO {SI,NO}
@attribute TIP_INSCRIPCION {TRANSFERENCIA_INTERNA,NORMAL,TRANSFERENCIA_EXTERNA,REINTEGRO,RESERVA_CUPO,EXO_TRANSF_INTERNA,EXO_NORMAL}
@attribute DESERTO {SI,NO}

@data
ADMINISTRACION_AMBIENTAL,Barranquilla_o_Municipio_del_Atlantico,NO,NO,2,PRIVADO,NO,MASCULINO,rango2,LABORA,LABORA,SI,SI,SI,NO,1,2,3,No
ADMINISTRACION_AMBIENTAL,Barranquilla_o_Municipio_del_Atlantico,NO,NO,3,PRIVADO,NO,FEMENINO,rango4,AMA_DE_CASA,LABORA,SI,SI,SI,NO,1,2,
ADMINISTRACION_AMBIENTAL,Fuera_del_Atlantico,SI,SI,3,PUBLICO,NO,FEMENINO,rango8,AMA_DE_CASA,LABORA,SI,SI,SI,NO,3,2,4,No_pertenece,NO,S
ADMINISTRACION_DE_EMPRESAS,Barranquilla_o_Municipio_del_Atlantico,SI,SI,1,PUBLICO,NO,MASCULINO,rango9,PENSIONADA,PENSIONADO,SI,SI,SI,S
ADMINISTRACION_DE_EMPRESAS,Barranquilla_o_Municipio_del_Atlantico,SI,SI,2,PUBLICO,NO,FEMENINO,rango6,AMA_DE_CASA,LABORA,SI,SI,SI,No_re
ADMINISTRACION_DE_EMPRESAS,Barranquilla_o_Municipio_del_Atlantico,SI,SI,1,PRIVADO,NO,MASCULINO,rango8,AMA_DE_CASA,LABORA,SI,SI,SI,SI,3
ADMINISTRACION_DE_EMPRESAS,Fuera_del_Pais,NO,NO,5,PRIVADO,NO,FEMENINO,rango4,LABORA,LABORA,SI,SI,SI,SI,1,1,3,No_pertenece,SI,NO,NO,550
ADMINISTRACION_DE_EMPRESAS,Fuera_del_Atlantico,NO,NO,3,PRIVADO,NO,MASCULINO,rango2,LABORA,LABORA,SI,SI,SI,SI,1,2,4,No_pertenece,SI,SI,
ADMINISTRACION_DE_EMPRESAS,Barranquilla_o_Municipio_del_Atlantico,NO,NO,3,PRIVADO,NO,FEMENINO,rango2,LABORA,LABORA,NO,NO,SI,NO,1,2,4,M
ADMINISTRACION_DE_EMPRESAS,Barranquilla_o_Municipio_del_Atlantico,NO,NO,2,PRIVADO,NO,MASCULINO,rango2,LABORA,LABORA,SI,SI,NO,NO,1,1,4,
```

Figura 27. Archivo ARFF

Fuente: Elaboración propia

Después de diligenciar los valores de las distintas características, puede observar la predicción arrojada por la aplicación en la Figura 28.

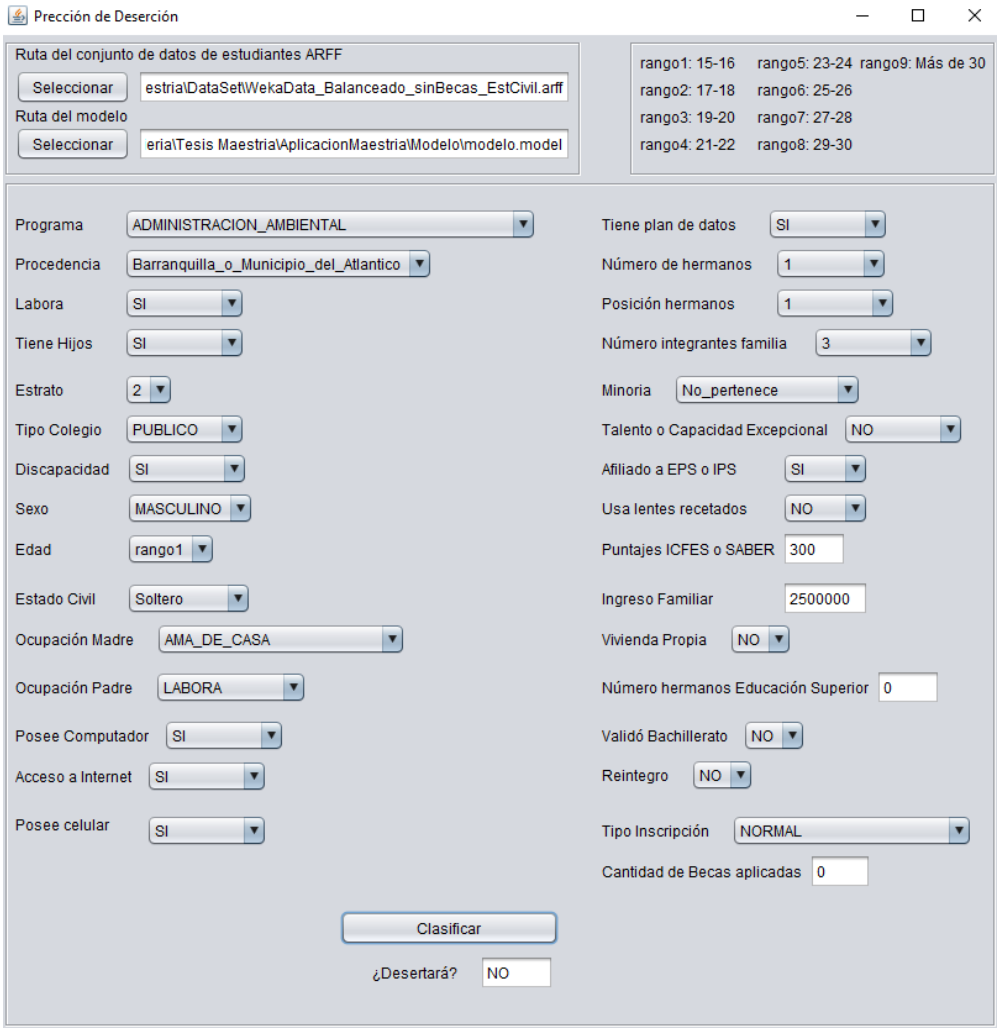


Figura 28. Formulario con los resultados arrojados por la aplicación

Fuente: Elaboración propia

Predicción masiva

Para facilitar la predicción masiva del aplicativo y ayudar a tomar decisiones de forma más ágil, se habilitó la carga masiva para la predicción de forma automática, subiendo un archivo Excel (xlsx), con treinta (30) características de la Tabla 8, excluyendo becas aplicadas y estado civil, agregando el ID en la primera columna. En la figura 29 parte derecha superior, se encuentra la opción para subir el archivo xlsx con los atributos de los estudiantes, para realizar la predicción masiva.

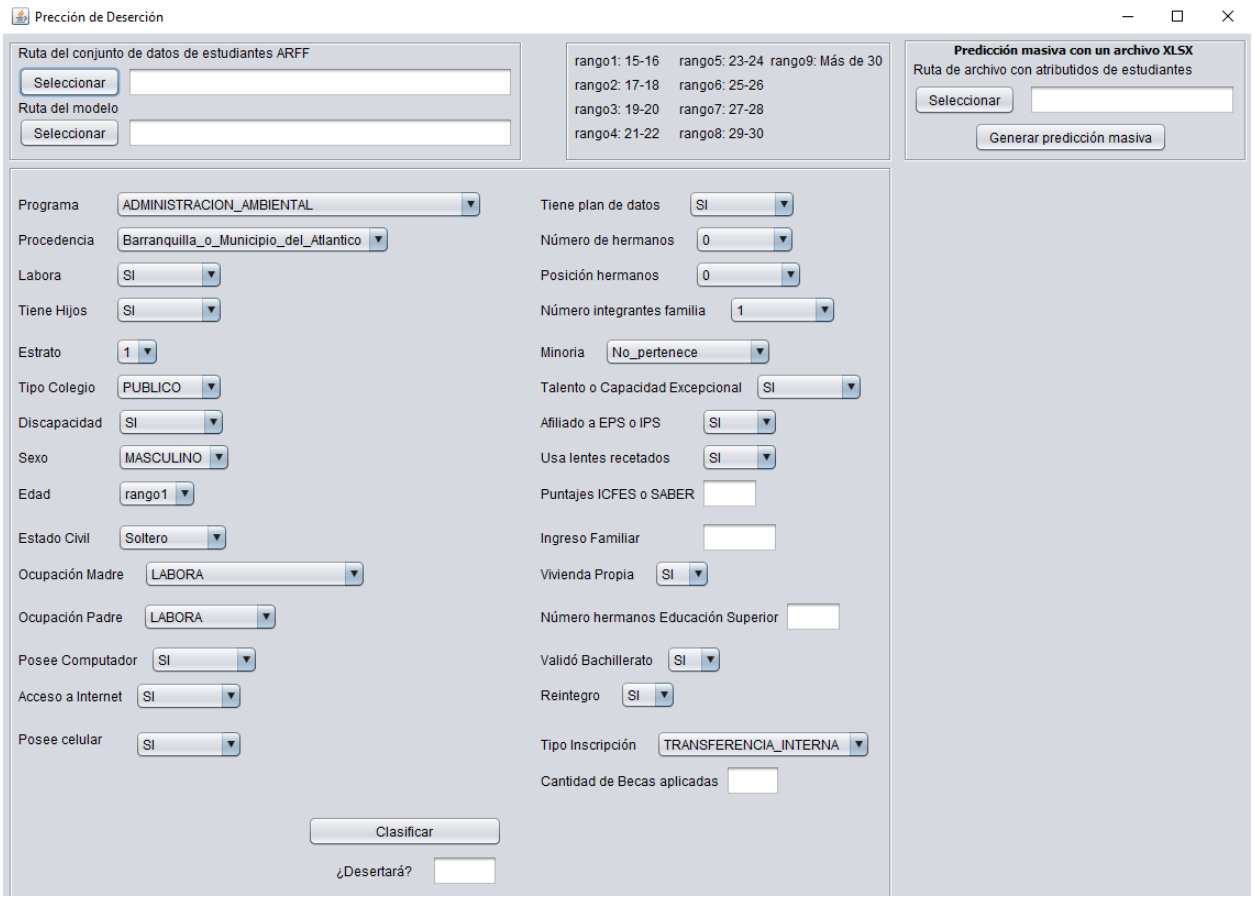


Figura 29. Predicción masiva

Fuente: Elaboración propia

Se debe seleccionar el archivo con los datos de los estudiantes a generar la predicción, no debe tener campos vacíos o nulos, no debe tener tildes, esto se debe hacer antes de subir el

archivo, de lo contrario, generará un error, impidiendo la ejecución de la predicción. Segundo, se selecciona el archivo ARFF y el modelo, se procede hacer clic en “Generar predicción masiva”, si todo se ejecuta correctamente, saldrá un mensaje indicando que la predicción se generó correctamente (Figura 30).

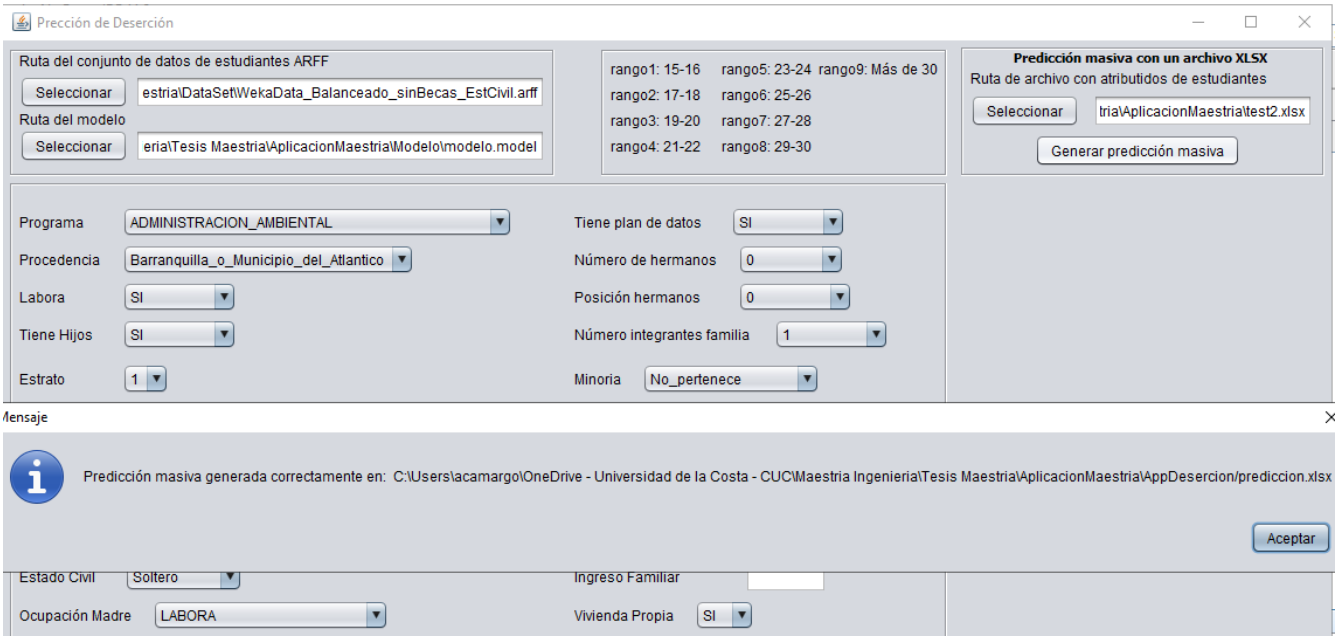


Figura 30. Ejecución predicción masiva

Fuente: Elaboración propia

Se genera un nuevo archivo de Excel adicionando la característica “desertó”, ver Figura 31, incluye “si” para los estudiantes que están en riesgo de desertar, y “no” para los estudiantes que no significan un riesgo de deserción. Es posible usar Excel (figura 32) u otra herramienta para el análisis de los datos, que al usuario final tomar la mejor decisión basado en los resultados.

Z	AA	AB	AC	AD	AE
VIVIENDA_PROPIA	NUM_HERMANOS_EDUC_SUPERIOR	VALIDO_BACHILLERATO	IND_REINTEGRO	TIP_INSCRIPCION	DESERTO
SI	1	NO	NO	TRANSFERENCIA_INTERNA	SI
NO	1	NO	NO	NORMAL	NO
NO	1	NO	NO	NORMAL	NO
SI	0	NO	NO	NORMAL	NO
NO	1	NO	NO	NORMAL	SI
SI	1	NO	NO	NORMAL	SI
NO	0	NO	NO	NORMAL	NO
NO	1	NO	NO	NORMAL	NO
NO	0	NO	NO	NORMAL	NO
NO	0	NO	NO	NORMAL	NO

Figura 31. Resultado predicción masiva

Fuente: Elaboración propia

Cuenta de IDENTIFICACION	Etiquetas de columna		
Etiquetas de fila	NO	SI	Total general
ADMINISTRACION_AMBIENTAL	2	1	3
123456		1	1
532323	1		1
555555	1		1
ADMINISTRACION_DE_EMPRESAS	5	2	7
1	1		1
2		1	1
3		1	1
4	1		1
5	1		1
6	1		1
7	1		1
Total general	7	3	10

Figura 32. Ejemplo tabla dinámica para análisis de resultado

Fuente: Elaboración propia

Mejoras futuras

Para mejorar futuras se plantean varias, permitiendo agilizar y acortando el proceso para la toma de decisiones.

- Integración con el sistema académico SICUC
- Ejecución automatizada
- Generación de alertas por medio de correo electrónico
- Entrenamiento del modelo con características académicas
- Integración con bases datos
- Integración con PowerBI para análisis visual intuitiva de los datos

Conclusiones

Después de analizar los resultados se determinó que el mejor algoritmo para clasificar la deserción estudiantil en la Universidad de la Costa CUC, es *RandomForest* con datos balanceados, la verificación de la efectividad de la técnica se obtuvo luego de realizar un proceso de prueba (test) mediante cross-validation utilizando igual 25 pliegues. La perspectiva ideal para construir el algoritmo es utilizar la información de todos los semestres en los que se matriculan los estudiantes, tomando como variable de clasificación aquella que define a los no desertores como aquellos estudiantes que terminaron su pregrado.

La predicción del algoritmo *RandomForrest* fue la mejor de las alternativas evaluadas (Maquinas de soporte vectorial y Redes Bayesianas). Del conjunto de datos disponible en el proyecto (1606), fue posible obtener los resultados, arrojando inicialmente una exactitud (*accuracy*) del 78.1% para el clasificador *ADTree*. Como los datos se encontraban desbalanceados, al aplicar SMOTE, los registros incrementaron a 2479, arrojando una exactitud de 84.1% con el clasificador *RandomForest*. Para lograr mejorar la predicción, se realizaron distintas simulaciones, estas consistieron en eliminar atributos y ver el impacto que estas tenían en el resultado final. Se encontró que los atributos Becas aplicadas y Estado Civil no tenían un impacto en la predicción, logrando mejorarla de 84.1% a 84.8%.

Cabe anotar que, al no utilizar todos los atributos, pudo influenciar en la obtención de una mejor predicción de la deserción. Así mismo, la poca cantidad de información obtenida para el proyecto de investigación. Entrenar el modelo con más datos, nos daría una mayor tasa de precisión de estudiantes en riesgo de desertar. Utilizar menos características incide en el tiempo de evaluación del modelo, lo cual en consecución puede ser útil si a futuro se plantea fortalecer el aplicativo con una interacción en tiempo real.

El modelo y la aplicación prototipo creados son un insumo importante para determinar los estudiantes en riesgo de desertar, ya que detecta 8 de 10 estudiantes en riesgo en los primeros semestres. La aplicación prototipo se usaría para detectar a los estudiantes de primer semestre, ayudando a tomar acciones correctivas por parte de Bienestar Universitario, disminuyendo el riesgo de deserción.

El siguiente paso para obtener un modelo con mayor precisión es incluir nuevas variables en el conjunto de datos, que han mostrado alguna evidencia de relación con la deserción en otros estudios realizados (Ver Tabla 6). Así mismo, entrenar los datos con la mayor cantidad de datos posibles.

Se propone generar a futuro un modelo de minería de datos, aplicando técnicas de machine learning, para realizar un proceso de predicción en tiempo real, entrenando el modelo con mayores cantidades de datos, aumentando la exactitud del mismo.

Referencias

Aboubakar, M., Kellil, M., Bouabdallah, A., & Roux, P. (2019). Toward Intelligent Reconfiguration of RPL Networks using Supervised Learning. *IFIP Wireless Days*, *2019-April*, 1–4. https://doi.org/10.1109/WD.2019.8734236

Aggarwal, C. (2015). *Data Mining: The Textbook*. Springer International. https://doi.org/10.1007/978-3-319-14142-8 ISBN

Ahuja, R., & Kankane, Y. (2017). Predicting the probability of student's degree completion by using different data mining techniques. *2017 Fourth International Conference on Image Information Processing (ICIIP)*, 1–4. https://doi.org/10.1109/ICIIP.2017.8313763

Alkhasawneh, R., & Hobson, R. (2011). Modeling student retention in science and engineering disciplines using neural networks. *2011 IEEE Global Engineering Education Conference, EDUCON 2011*, 660–663. https://doi.org/10.1109/EDUCON.2011.5773209

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Askinadze, A., & Conrad, S. (2017). Application of the Dynamic Time Warping Distance for the Student Drop-out Prediction on Time Series Data. *Proceedings of the 10th International Conference on Educational Data Mining*, 342–343.

Azevedo, A., & Santos, M. F. (2008). KDD, semma and CRISP-DM: A parallel overview. *MCCSIS'08 - IADIS Multi Conference on Computer Science and Information Systems; Proceedings of Informatics 2008 and Data Mining 2008, June*, 182–185.

Aziz, A. A., Ismail, N. H., Ahmad, F., & Hassan, H. (2015). A framework for students' academic

performance analysis using naïve bayes classifier. *Jurnal Teknologi*, *75*(3), 13–19. https://doi.org/10.11113/jt.v75.5037

Barbosa Manhães, L. M., Serra da Cruz, S. M., & Zimbrão, G. (2014). WAVE: an Architecture for Predicting Dropout in Undergraduate Courses using EDM. *Proceeding SAC '14 Proceedings of the 29th Annual ACM Symposium on Applied Computing*, 243–247. https://doi.org/10.1145/2554850.2555135

Barker, K., Trafalis, T., & Reed Rhoads, T. (2004). LEARNING FROM STUDENT DATA. *Proceedings of the 2004 Systems and Information Engineering Design Symposium Matthew*. https://doi.org/10.1109/SIEDS.2004.239819

Barnes, T., Desmarais, M., Romero, C., & Ventura, S. (2009). EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining. In *EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining*.

Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T., & Popelínský, L. (2012). Predicting drop-out from social behaviour of students. *Proceedings of the 5th International Conference on Educational Data Mining*, *Dm*, 103–109.

Beaulac, C., & Rosenthal, J. S. (2019). *Predicting University Students ' Academic Success and Choice of Major using Random Forests*.

Beltran, B. (2016). *MINERÍA DE DATOS* (Vol. 30, Issue 1). https://doi.org/10.1016/0032-0633(82)90071-X

Betancourt, G. A. (2005). LAS MÁQUINAS DE SOPORTE VECTORIAL (SVMs). *Scientia Et Technica*, *XI*(27), 67–72. https://doi.org/10.22517/23447214.6895

Birjali, M., Beni-hssane, A., & Erritali, M. (2018). *Learning with Big Data Technology: The*

Future of Education. *565*. https://doi.org/10.1007/978-3-319-60834-1

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, 144–152. https://doi.org/10.1145/130385.130401

Brown, M. S. (2014). *Data Mining for Dummies*. https://doi.org/10.1007/978-1-4614-7669-6

Burgueño, M. J., García-Bastos, J. L., & González-Buitrago, J. M. (1995). ROC curves in the evaluation of diagnostic tests. *Medicina Clínica*, *104*(17), 661–670.

Cambruzzi, W., Rigo, S. J., & Barbosa, J. L. V. (2015). Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach. *Journal of Universal Computer Science*, *21*(1), 23–47.

Carmona Suárez, E. J. (2014). Máquinas de Vectores Soporte (SVM). *Dpto. de Inteligencia Artificial, ETS de Ingeniería Inforática, Universidad Nacional de Educación a Distancia (UNED)*, 1–25. http://www.ia.uned.es/~ejcarmona/publicaciones/[2013-Carmona] SVM.pdf

Castaño, E., Gallón, S., Gómez, K., & Vásquez, J. (2004). Deserción estudiantil universitaria una aplicación de modelos de duración. *Lecturas de Economia*, *60*(60), 39–65.

Chai, K. E. K., & Gibson, D. (2015). Predicting the risk of attrition for undergraduate students with time based modelling. *Proceedings of the 12th International Conference on Cognition and Exploratory Learning in the Digital Age, CELDA 2015*, *Celda*, 109–116.

Cheewaprakobkit, P. (2013). Study of factors analysis affecting academic achievement of undergraduate students in international program. *Lecture Notes in Engineering and Computer Science*, *2202*, 332–336.

Christian, T. M., & Ayub, M. (2014). Exploration of classification using NBTree for predicting students' performance. *Proceedings of 2014 International Conference on Data and*

Software Engineering, ICODSE 2014, 1–6. https://doi.org/10.1109/ICODSE.2014.7062654

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Mach. Learn. , 20 (3)*, *44*(13), 273–297.

Predicting Students Drop Out A Case Study, (2009).

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory and Practice*, *13*(1), 17–35. https://doi.org/10.2190/CS.13.1.b

Devasia, T., Vinushree, T. P., & Hegde, V. (2016). Prediction of students performance using Educational Data Mining. *Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016*, 91–95. https://doi.org/10.1109/SAPIENCE.2016.7684167

Dharmawan, T., Ginardi, H., & Munif, A. (2018). Dropout Detection Using Non-Academic Data. *Proceedings - 2018 4th International Conference on Science and Technology, ICST 2018*, *1*, 1–4. https://doi.org/10.1109/ICSTC.2018.8528619

Edwards, W., & Fasolo, B. (2001). *Decision Technology*. 581–606.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. Van. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*(August 2017), 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). *Bayesian Network Classifier*. 131–163.

Gandhi, R. (2018). *Support Vector Machine - Introduction to Machine Learning Algorithms*. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

García, J. G., Puga, J. L., Cano Guillén, C. J., Gea, A. B., & de la Fuente Sánchez, L. (2006). Aplicación de las redes bayesianas al modelado de las actitudes emprendedoras. *IV Congreso de Metodología de Encuestas, August 2015*, 235–242.

Gulati, H. (2015). Predictive analytics using data mining technique. *2015 International Conference on Computing for Sustainable Global Development, INDIACom 2015*, 713–716.

Güner, N., Yaldir, A., Gündüz, G., Çomak, E., Tokat, S., & Iplikçi, S. (2014). Predicting academically at-risk engineering students: A soft computing application. *Acta Polytechnica Hungarica*, *11*(5), 199–216. https://doi.org/10.12700/aph.11.05.2014.05.12

Guzmán Ruiz, C., Muriel Durán, D., & Franco Gallego, J. (2009). *Deserción estudiantil en la educación superior colombiana. Metodología de seguimiento, diagnóstico y elementos para su prevención*. http://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_libro_desercion.pdf

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques*. https://doi.org/10.1109/ICMIRA.2013.45

Hasbun, T., Araya, A., & Villalon, J. (2016). Extracurricular activities as dropout prediction factors in higher education using decision trees. *Proceedings - IEEE 16th International Conference on Advanced Learning Technologies, ICALT 2016*, 242–244. https://doi.org/10.1109/ICALT.2016.66

Heredia, D., Amaya, Y., & Barrientos, E. (2015). Student Dropout Predictive Model Using Data Mining Techniques. *IEEE Latin America Transactions*, *13*(9), 3127–3134.

https://doi.org/10.1109/TLA.2015.7350068

Hernandez Gonzalez, A. G., Melendez Armenta, R. A., Morales Rosales, L. A., Garcia Barrientos, A., Tecpanecatl Xihuitl, J. L., & Algredo, I. (2016). Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-Mexico. *IEEE Latin America Transactions*, *14*(11), 4573–4578. https://doi.org/10.1109/TLA.2016.7795831

Hoffait, A. S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*, 1–11. https://doi.org/10.1016/j.dss.2017.05.003

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model. *Procedia CIRP*, *79*, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

Jin, Q., Imbrie, P. K., Lin, J. J. J., & Chen, X. (2011). A multi-outcome hybrid model for predicting student success in engineering. *ASEE Annual Conference and Exposition, Conference Proceedings*.

Kabakchieva, D., Stefanova, K., & Kisimov, V. (2011). Analyzing university data for determining student profiles and predicting performance. *EDM 2011 - Proceedings of the 4th International Conference on Educational Data Mining*.

Kalles, D., & Pierrakeas, C. (2006a). Analyzing student performance in distance learning with genetic algorithms and decision trees. *Applied Artificial Intelligence*, *20*(8), 655–674. https://doi.org/10.1080/08839510600844946

Kalles, D., & Pierrakeas, C. (2006b). Using genetic algorithms and decision trees for a posteriori analysis and evaluation of tutoring practices based on student failure models. *IFIP International Federation for Information Processing*, *204*(August), 9–18. https://doi.org/10.1007/0-387-34224-9_2

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees. *Nat Biotechnol*, *23*(1), 1–7. https://doi.org/10.1038/nbt0908-1011

Kotsiantis, S. B., & Pintelas, P. E. (2005). Predicting students' marks in Hellenic Open University. *Proceedings - 5th IEEE International Conference on Advanced Learning Technologies, ICALT 2005*, *2005*, 664–668. https://doi.org/10.1109/ICALT.2005.223

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, *18*(5), 411–426. https://doi.org/10.1080/08839510490442058

Krishna Kishore, K. V., Venkatramaphanikumar, S., & Alekhya, S. (2014). Prediction of student academic progression: A case study on Vignan University. *2014 International Conference on Computer Communication and Informatics: Ushering in Technologies of Tomorrow, Today, ICCCI 2014*, *2*, 1–6. https://doi.org/10.1109/ICCCI.2014.6921731

Kumar Baradwaj, B., & Pal, S. (2011). Mining Educational Data to Analyze Students' Performance. *JACSA) International Journal of Advanced Computer Science and Applications*, *02*.

Lee, S., & Chung, J. Y. (2019). The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. *Applied Sciences*, *9*(15), 3093. https://doi.org/10.3390/app9153093

Lesinski, G., Corns, S., & Dagli, C. (2016). Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy. *Procedia Computer Science*, *95*, 375–382. https://doi.org/10.1016/j.procs.2016.09.348

López de Ullibarri, G. I., & Píta Fernández, S. (1998). Curvas ROC. *Cad Aten Primaria*, *5*(4), 229–235.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers and Education*, *53*(3), 950–965. https://doi.org/10.1016/j.compedu.2009.05.010

Manhães, L. M. B., Da Cruz, S. M. S., & Zimbrão, G. (2014). The impact of high dropout rates in a large public brazilian university a quantitative approach using educational data mining. *CSEDU 2014 - Proceedings of the 6th International Conference on Computer Supported Education*, *3*, 124–129. https://doi.org/10.5220/0004958601240129

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Márquez-Vera, C., Romero Morales, C., & Ventura Soto, S. (2013). Predicting school failure and dropout by using data mining techniques. *Revista Iberoamericana de Tecnologias Del Aprendizaje*, *8*(1), 7–14. https://doi.org/10.1109/RITA.2013.2244695

Mayilvaganan, M., & Kalpanadevi, D. (2015). Comparison of classification techniques for predicting the performance of students academic environment. *2014 International Conference on Communication and Network Technologies, ICCNT 2014*, *2015-March*, 113–118. https://doi.org/10.1109/CNT.2014.7062736

Ministerio de Educación de Colombia. (2006). La Revolución Educativa 2002 – 2006. *Media*, 1–6.

Ministerio de Educación de Colombia. (2019). *Qué es el SPADIES*. https://www.mineducacion.gov.co/sistemasinfo/spadies/Informacion-Institucional/254648:Que-es-el-SPADIES

Miranda, M. A., & Guzmán, J. (2017). Análisis de la deserción de estudiantes universitarios

usando técnicas de minería de datos. *Formacion Universitaria*, *10*(3), 61–68. https://doi.org/10.4067/S0718-50062017000300007

Mishra, A. (2018). *Metrics to Evaluate your Machine Learning Algorithm*. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Mishra, T., Kumar, D., & Gupta, S. (2014). Mining students' data for prediction performance. *International Conference on Advanced Computing and Communication Technologies, ACCT*, 255–262. https://doi.org/10.1109/ACCT.2014.105

Mitchell, T. M. (1997). Machine Learning. In *McGraw-Hill Science/Engineering/Math*.

Moseley, L. G., & Mead, D. M. (2008). Predicting who will drop out of nursing courses: A machine learning exercise. *Nurse Education Today*, *28*(4), 469–475. https://doi.org/10.1016/j.nedt.2007.07.012

Mustafa, M. N., Chowdhury, L., & Kamal, M. S. (2012). Students dropout prediction for intelligent system from tertiary level in developing country. *2012 International Conference on Informatics, Electronics and Vision, ICIEV 2012*, 113–118. https://doi.org/10.1109/ICIEV.2012.6317441

Oskouei, R. J., & Askari, M. (2014). Predicting Academic Performance with Applying Data Mining Techniques (Generalizing the results of two Different Case Studies). *Computer Engineering and Applications Journal*, *3*(2), 79–88. https://doi.org/10.18495/comengapp.v3i2.81

Osmanbegovi, E. (2012). Data Mining Approach for Predicting Student Performance. *Economic Review : Journal of Economics and Business*, *X*(1), 3–12.

Peralta, B., Poblete, T., & Caro, L. (2017). Automatic feature selection for desertion and

graduation prediction: A chilean case. *Proceedings - International Conference of the Chilean Computer Science Society, SCCC*. https://doi.org/10.1109/SCCC.2016.7836055

Pereira, R. T., Romero, A. C., & Toledo, J. J. (2013). Extraction student dropout patterns with data mining techniques in undergraduate programs. *IC3K 2013; KDIR 2013 - 5th International Conference on Knowledge Discovery and Information Retrieval and KMIS 2013 - 5th International Conference on Knowledge Management and Information Sharing, Proc.*, 136–142. https://doi.org/10.5220/0004543001360142

Pérez, A., Grandón, E. E., Caniupán, M., & Vargas, G. (2019). Comparative Analysis of Prediction Techniques to Determine Student Dropout: Logistic Regression vs Decision Trees. *Proceedings - International Conference of the Chilean Computer Science Society, SCCC*, *2018-Novem*. https://doi.org/10.1109/SCCC.2018.8705262

Perez, B., Castellanos, C., & Correal, D. (2018). Applying Data Mining Techniques to Predict Student Dropout: A Case Study. *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence, ColCACI 2018 - Proceedings*, 1–6. https://doi.org/10.1109/ColCACI.2018.8484847

Perez, M. (2014). *Minería de datos a treves de ejemplos*. 22. http://www.rclibros.es/pdf/capitulo_mineria.pdf

Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy, D., & Strohecker, C. (2004). Affective learning - a manifesto. *BT Technology Journal*, *22*(4), 253–269. https://doi.org/10.1023/B:BTTJ.0000047603.37042.33

Pradeep, A., Das, S., & Kizhekkethottam, J. J. (2015). Students dropout factor prediction using EDM techniques. *Proceedings of the IEEE International Conference on Soft-Computing*

and Network Security, ICSNS 2015, 1–7. https://doi.org/10.1109/ICSNS.2015.7292372

Quadri, M., & Kalyankar, D. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer*, *10*(2), 2–5. http://computerresearch.org/stpr/index.php/gjcst/article/viewArticle/128

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/bf00116251

Recuero, P. (2018). *Machine Learning a tu alcance: La matriz de confusión*. https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/

Salazar, A., Gosálbez, J., Bosch, I., Miralles, R., & Vergara, L. (2004). A case study of knowledge discovery on academic achievement, student desertion and student retention. *ITRE 2004 - 2nd International Conference on Information Technology: Research and Education - Proceedings*, *January*, 150–154. https://doi.org/10.1109/itre.2004.1393665

Sangodiah, A., Beleya, P., Muniandy, M., Heng, L. E., & Ramendran Spr, C. (2015). Minimizing student attrition in higher learning institutions in Malaysia using support vector machine. *Journal of Theoretical and Applied Information Technology*, *71*(3), 377–385.

Santana, M. A., Costa, E. B., Neto, B. F. S., Silva, I. C. L., & Rego, J. B. A. (2015). A predictive model for identifying students with dropout profiles in online courses. *CEUR Workshop Proceedings*, *1446*.

Şara, N. B., Halland, R., Igel, C., & Alstrup, S. (2015). High-school dropout prediction using machine learning: A Danish large-scale study. *23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015 - Proceedings*, *April*, 319–324.

Saravanan, R., & Sujatha, P. (2018). Algorithms : A Perspective of Supervised Learning

Approaches in Data Classification. *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Iciccs*, 945–949.

Sarker, F., Tiropanis, T., & Davis, H. C. (2014). Linked data, data mining and external open data for better prediction of at-risk students. *Proceedings - 2014 International Conference on Control, Decision and Information Technologies, CoDIT 2014*, 652–657. https://doi.org/10.1109/CoDIT.2014.6996973

Segura-Morales, M., & Loza-Aguirre, E. (2018). Using Decision Trees for Predicting Academic Performance Based on Socio-Economic Factors. *Proceedings - 2017 International Conference on Computational Science and Computational Intelligence, CSCI 2017*, 1132–1136. https://doi.org/10.1109/CSCI.2017.197

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*(February 2016), 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Sharabiani, A., Karim, F., Sharabiani, A., Atanasov, M., & Darabi, H. (2014). An enhanced bayesian network model for prediction of students' academic performance in engineering programs. *IEEE Global Engineering Education Conference, EDUCON, April*, 832–837. https://doi.org/10.1109/EDUCON.2014.6826192

Siri, A. (2015). Predicting Students' Dropout at University Using Artificial Neural Networks. *Italian Journal of Sociology of Education*, *7*(2), 225–247.

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to Predict Dropout in University Students with Machine Learning. *2018 IEEE International Work Conference on Bioinspired Intelligence, IWOBI 2018 - Proceedings*. https://doi.org/10.1109/IWOBI.2018.8464191

Tair, M. M. A. (2015). *Mining Educational Data to Improve Students ' Performance : A Case Study Mining Educational Data t o Improve Students ' Performance : A Case Study*. *October*.

Thomas, E. H., & Galambos, N. (2004). What satisfies students? Mining student-opinion data with regression and decision tree analysis. *Research in Higher Education*, *45*(3), 251–269.

Timarán Pereira, S. R., Hernández Arteaga, I., Caicedo Zambrano, S. J., Hidalgo Troya, A., & Alvarado Pérez, J. C. (2016). Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. *Descubrimiento de Patrones de Desempeño Académico Con Árboles de Decisión En Las Competencias Genéricas de La Formación Profesional*, *2016*, 63–86. https://doi.org/10.16925/9789587600490

Tsai, C. F., Tsai, C. T., Hung, C. S., & Hwang, P. Sen. (2011). Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation. *Australasian Journal of Educational Technology*, *27*(3), 481–498. https://doi.org/10.14742/ajet.956

Universidad Pedagógica y Tecnológica de Colombia. (2004). *Unidad 1 Estadistica Descriptiva*. https://virtual.uptc.edu.co/ova/estadistica/docs/libros/h_men_prob_est/lecciones_html/un1/1_8_3.html

Veitch, W. R. (2004). Identifying Characteristics of High School Dropouts: Data Mining with A Decision Tree Model. *Online Submission*, 1–11.

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, *24959*, 29–39. https://doi.org/10.1.1.198.5133

Yehuala, M. A. (2015). Application Of Data Mining Techniques For Student Success And Failure Prediction The Case Of DebreMarkos University. *International Journal of Scientific & Technology Research*, *4*(4), 91–94.

Zaki, M., & Meira, W. J. (2013). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. https://doi.org/10.1145/3054925

Zeng, W., Chin, S.-C., Zeimet, B., Kuang, R., & Chi, C.-L. (2017). Dropout Prediction in Home Care Training. *Proceedings of the 10th International Conference on Educational Data Mining*, 442–447.

Zhang, Y., & Oussena, S. (2010). USE DATA MINING TO IMPROVE STUDENT RETENTION IN HIGHER EDUCATION – A CASE STUDY. *Middlesex University Research Repository*.